Screening with FAST earlyReading measures: An examination of individual measures and composite scores

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Barbara D. Monaghen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

June 2014

Abstract

This purpose of this project was to examine the validity of FAST Early Reading (earlyReading™) for use as universal screeners in kindergarten and first grade. Specifically this study examined and compared the concurrent and predictive validity and the diagnostic accuracy of single (N = 12) and combined measures. earlyReading™ measures was administered to kindergarten (N=223) and first grade students (N=180) in the fall, winter and spring of one school year. The Group Reading Assessment and Diagnostic Evaluation (GRADE) was administered at the end of the school year and used as the criterion measure. Study 1 evaluated the concurrent and predictive validity of individual measures. Across all time points in kindergarten it was determined that screening batteries of three to four measures had comparable concurrent and predictive validity to the full screening battery consisting of six to eight measures. In first grade, screening batteries of two to three measures in the fall, and one to two measures in the winter and spring were found comparable to the full screening battery consisting of five to six measures. Study 2 evaluated the diagnostic accuracy of single and composite scores was examined. In kindergarten, minimal levels of diagnostic accuracy were met using composites with two measures in the fall and winter, but not spring. By winter and spring of first grade, use of a single screening measure exceeded minimum standards of diagnostic accuracy for screening and was comparable to larger screening batteries. Results suggest that earlyReading was efficient and technically adequate for tri-annual universal screening in kindergarten and first grade. Implications and future research are discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1

The 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA) provided states with the option to "use a process that determines if a child responds to scientific, research based intervention" as an alternative to the severe discrepancy approach for prevention and identification of specific learning disability (SLD). Prior to this reauthorization, dissatisfaction was expressed regarding the severe discrepancy approach, often described as the "wait to fail" model (Berkeley, Bender, Gregg-Peaster, & Saunders, 2009; Bradley, Danielson, & Doolittle, 2007). There was concern that students were identified too late, that inconsistent identification occurred, and that prior models of disability determination lacked instructional utility (NJCLD, 2007). Response to intervention (RTI) quickly emerged as a viable alternative.

RTI is a framework through which schools provide early evidence-based intervention to students at risk of academic and behavioral difficulties. Through this framework, students who respond to intervention continue with the general education curriculum. Those that do not respond may experience intervention changes or a special education referral. Within these models, students become eligible for special education based on a dual discrepancy in rate and level of achievement (i.e., underachievement and insufficient response to intervention are required; Fuchs, Fuchs, McMaster, & Al Otaiba, 2003; Swanson, Harris, & Graham, 2003). With the IDEA regulations in place, over 70% of states were implementing RTI by 2009, with the remaining 30% in the process of implementing some form of RTI by 2008 (Hoover, Baca, Wexler-Love, & Saenz, 2008).

With wide-spread adoption and no universal set of procedures to implement RTI, variation across implemented models was found (Zirkel & Thomas, 2010). Despite these

differences, five core components are reiterated throughout the literature on RTI. These include high quality research-based instruction in general education, universal screening for academic and behavior problems, continuous progress monitoring, multiple tiers of progressively more intense instructions and intervention, and use of fidelity measures (Gessler-Werts, Lambert, & Carpenter, 2009; Zirkel & Thomas, 2010). These components have extended the focus of RTI beyond SLD identification and toward early identification and intervention for students at risk of academic and behavioral difficulties. While the goal of RTI is early, accurate, and consistent identification with strong links to instruction in the classroom, unresolved issues are apparent with the reality of RTI implementation.

The present study focused on the RTI core component of universal screening for reading among kindergarten and first grade students. Universal screening is emerging as a critical tool within RTI because it facilitates the systematic assessment of reading for all students and, thereby, helps identify those in need of tier II intervention and/or more in depth diagnostic assessment. Where tier I is generally defined as the general curriculum, tier II services provide students with extra support usually in the form of short-term targeted interventions. Characteristics of universal screening include time efficient, technically adequate, cost effective sampling of reading skills which can be easily administered, scored and interpreted to inform instruction and predict future performance. Universal screening is typically administered up-to-three times per year to all students (Ikeda, Neessen, & Witt, 2008).

While universal screening is purported as essential to the success of RTI implementation (Hoover & Love, 2011), not surprisingly there is variability of use across

states and districts. An examination of universal screening revealed that only 77% of surveyed general and special educators used screening to identify students for intervention and only 23% of those educators indicated that screening was *always* used (Martinez & Young, 2011). Moreover, similar to the severe discrepancy model, dissatisfaction was expressed with the validity of universal screening in identifying at-risk students in reading for kindergarten and first grade students (Denton, 2012; D. Fuchs & Fuchs, 2006; D. Fuchs, Fuchs, & Compton, 2012; Hughes & Dexter, 2011; Reynolds & Shaywitz, 2009).

Different approaches to universal screening exist with variation in prediction accuracy observed across approaches. The most common and straightforward approach to universal screening involves the administration of one screening measure across three time points during the school year. Students who score below a pre-specified cut-point are considered for intervention services. This method is referred to as the Direct Route (DR) method of universal screening (Jenkins, Hudson, & Johnson, 2007). While it is efficient and cost effective, the technical adequacy of the DR approach is questioned. The alternative approach called the progress monitoring (PM) approach, uses 5 to 8 weeks of progress monitoring to follow up after screening to determine sufficient (or insufficient) rate of improvement in addition to student's initial level (Compton, Fuchs, Fuchs, & Bryant, 2006). Progress monitoring after initial screening improves accuracy of identification at the cost of delayed intervention for at-risk students. Although the PM approach to universal screening has produced more acceptable levels of technical adequacy, the additional weeks of progress monitoring sacrifices the efficiency and cost effectiveness achieved in the DR approach.

The current project examined the technical adequacy the DR approach with multiple measure screening batteries. With most schools reportedly use the DR approach for identification of intervention services (Mellard, Byrd, Johnson, Tollefson, & Boesche, 2004), test validation of universal screening of early reading is timely and important. Chapter 2 is a review of the literature on early reading and universal screeners. Chapter 3 presents Study 1, which examines the concurrent and predictive validity of using single and multiple reading measures for universal screening in kindergarten and first grade. Chapter 4 follows with Study 2, which examines the diagnostic accuracy of the same single and multiple reading measures. Chapter 5 ends with an integrative discussion of the findings and implications.

Chapter 2

Early reading measures for use in universal screening are rooted in approaches to reading instruction. Throughout the twentieth century, educators have debated the use of whole-language (or meaning-emphasis), code-emphasis, or a balanced approach to reading instruction. Chall (1996) describes the controversial study of reading as "The Great Debate." With plenty of opinions and empirical evidence claiming to provide support for different instructional approaches, the National Reading Panel (NRP) was charged by Congress to examine the status of research-based knowledge and the effectiveness of instruction related to reading (National Reading Panel, 2000). The NRP concluded that the best approach to reading instruction included explicit instruction in phonological awareness, alphabetic principle, fluency with connected text, vocabulary and comprehension. The NRP majority report found that phonemic awareness and phonics instruction significantly outperformed alternative forms of training, such as wholistic and meaning-centered (National Reading Panel, 2000). The NRP recognized, however, that instructional methods are often not used in isolation. That is, a teacher might employ both phonemic awareness and wholistic instruction within the classroom. Moreover, the NRP recognized that implementation of an instructional method, such as phonics, may differ across classrooms and specific strategies used. Although the effectiveness of different instructional methods is outside the scope of this review, phonemic awareness and phonics instruction are introduced only to highlight an example of different instructional methods in early reading.

Phonemic awareness instruction includes a variety of activities designed to promote the skills and abilities in individuals to identify and manipulate phonemes in a

spoken word. Activities can include phoneme isolation, identification, categorization, blending, segmenting, deletion, addition, and substitution (Stahl & McKenna, 2001). These tasks require students to identify and manipulate individual sounds, or words that typically consist of two to four phonemes. Phonics instruction includes teaching the alphabetic system which includes phoneme-grapheme correspondence and basic spelling patterns (Adams, 1994). Phonics instruction is distinguished from phonemic awareness instruction as it promotes the understanding of the relationship between letters and sounds in written language; unlike phonemic awareness instruction, which emphasizes only the sounds in spoken language (Stahl & McKenna, 2001). Phonics and phonemic awareness instruction can be delivered in a number of ways; however, systematic and explicit instruction produces the greatest impact on children's reading achievement (NRP, 2000).

Models of Reading Development

Differing approaches to reading instruction are grounded in one or more of the major reading theories. Although reading theory dates back to at least 1925 with William Gray's stages of reading development (Chall, 1983; Indrisano & Chall, 1995), the most elaborate stage model focused largely on the transitions between stages and took into account the influence of different methods of reading instruction (Chall, 1983). Chall's model begins with Stage 0 where children from birth to approximately age six learn the concepts of reading and start to read stories based on pictures. Transition into Stage 1 is indicated by insight into the alphabetic principle, not immediate word recognition. In Stages 1 (decoding) and 2 (fluency), children progress from learning the alphabetic principle to reading simple texts with high frequency words, to acquiring fluency and

automaticity in reading familiar texts. At the end of Stage 2, reading moves from the oral mode of Stage 1, to silent reading by 3rd grade. By Stage 3, children transition from learning to read, to reading to learn and encounter increasingly unfamiliar material. Stages 4 and 5 are often reached in high school and college when texts are varied and complex in content, and require critical thinking skills to understand and learn from.

Other popular models include Laberge and Samuels (1974) theory of automatic information processing in reading, Ehri's (1999, 2005) "phase" theory of reading development, and Gough and Tunmer's (1986) and Hoover and Gough's (1990) "Simple View of Reading," (SVR). In Laberge and Samuels' theory, learning is evaluated on the basis of accuracy and automaticity, where accuracy must be achieved before automaticity. In Ehri's phase theory, learners are characterized into one of four sequential phases to include Pre-alphabetic, Partial Alphabetic, Full Alphabetic and Consolidated. Finally, within the SVR the product of decoding (D) and comprehension (C) equals reading (R; R = D x C). Ouellette and Beers (2010) support an even more complex model of reading, suggesting the components of reading change throughout development. In their study, different components of reading were found to influence reading comprehension across grade levels.

Despite differences in major theories of reading development, it is clear that the components of early reading are distinct from that of more advanced reading. Closer examination of theories highlight the influence of environmental and context clues, alphabetic principle, phonological awareness, and language proficiency for beginning readers. With greater similarities than differences across theories, the question becomes how best to measure the components of early reading most efficiently.

Indicators of Early Reading

Increasingly over the last 20 years, substantial efforts to create reading assessments that align with reading theory and instructional practices are observed in the literature. In addition to assessments that are explicitly linked to instruction, is the need for reading assessments to be valid, reliable and efficient for use in the classroom. Curriculum-Based Measurement (CBM) has emerged as a key tool within these efforts. CBM is a set of standardized procedures that are performance based, easy to administer and interpret scores within and across individuals. CBM can be repeatedly administered over short intervals, and uses materials that are representative and equivalent to those used in instruction (Deno, 2003). CBM procedures are used in practice with a number of academic skill areas, but especially with reading. Curriculum-Based Measurement of oral Reading (CBM-R) is a general outcome measure (GOM), as it is designed to measure the more general behavior of reading (L. S. Fuchs & Deno, 1991). The procedures require students to read from a grade or instructional level passage for one minute while words read correctly are recorded. Errors are documented as an omission, reversal, or miscue. Additionally, if the student does not read a word in 3-seconds, the word is provided to the student and counted as an error. Using this procedure, students earn a score for both accuracy and rate (automaticity). Accuracy is defined as the percentage of words read correctly, where rate is defined as the number of words read correctly in one minute.

The link between reading fluency and comprehension is well established in the literature (L. S. Fuchs, Fuchs, & Maxwell, 1988; Kranzler, Brownell, & Miller, 1998; Markell & Deno, 1997), in addition to validation of the relationship between oral reading fluency and reading theory (Shinn, Good, Knutson, TIlly, & Collins, 1992).

Nevertheless, the nature of the magnitude of this relationship across the development of reading has been met with criticism in the literature. In the earlier grades, especially kindergarten and beginning first grade, floor effects are observed with traditional CBM-R procedures (L. S. Fuchs, Fuchs, & Compton, 2004). These floor effects for assessments using connected text are consistent with the unique components of early reading identified by reading theory. For example, students who are acquiring skills related to the alphabetic principle and phonemic awareness may not show growth on assessment using connected text; however, growth in these early reading skills may be demonstrated using alternate assessments that are predictive of later reading achievement. Several groups have extended CBM procedures to include basic literacy. Broad descriptions of the major assessment packages are reviewed followed by a more detailed review of the psychometric evidence of each indicator of early reading in the section that follows.

Dynamic Indicators of Basic Skills (DIBELS). In 1996, research validating the use of DIBELS measures first appeared in the literature with initial publication of commercially available materials in 2002 (Good & Kaminski, 1996; Good & Kaminski, 2002a). Consistent with reading theory, DIBELS used measures that broke reading into components, namely phonological awareness and language development (Good & Kaminski, 1996). DIBELS Next (Good et al., 2011), the most recent edition of DIBELS, further broke these components into the five categories consistent with the essential skills of reading as reported by the National Reading Panel (2000). CBM procedures were used to identify students who are at-risk of developing reading difficulties in kindergarten through sixth grade using seven measures (Kaminski, Cummings, Powell-Smith, & Good, 2008). The measures include DIBELS Initial Sound Fluency (ISF), now called

First Sound Fluency (FSF), and Phonemic Segmentation Fluency (PSF) which assess phonemic awareness. DIBELS Nonsense Word Fluency (NWF) is the sole measure of alphabetic principal. The fourth measure, DIBELS Letter Naming Fluency (LNF) is not directly linked to reading (Adams, 1994; Kaminski, et al., 2008); however, the measure is predictive of later reading achievement. The last two measures, Word Use Fluency (WUF) is a measure of vocabulary and oral language, and Retell Fluency (RTF) is a measure of and reading comprehension.

AIMSweb and easyCBM. AIMSweb and easyCBM published a similar set of early literacy indicators around the same time (Alonzo & Tindal, 2004; M. M. Shinn & Shinn, 2002). AIMSweb indicators include LNF, PSF, and NWF in addition to Letter Sound Fluency (LSF). One other distinction between AIMSweb and DIBELSNext is the designation of the early literacy measures as GOMs or Specific Subskill Mastery Measurement (SSMM; Fuchs & Deno, 1991). DIBELS Next highlights all DIBELS measures as "economical and efficient indicators of a student's progress toward achieving a general outcome such as reading or phonemic awareness" (Good et al., 2011, pg. 5). In contrast, early literacy indicators are viewed as SSMM within the AIMSweb framework. That is, the AIMSweb early literacy measures are considered to function best within a mastery learning model where a short-term measurement approach is taken. easyCBM measures include LNF, LSF, PSF, and a high frequency word reading task.

FAST earlyReading. More recently, a set of early literacy indicators called earlyReading, were developed through FAST (Formative Assessment System for Teachers; formally known as the Formative Assessment Instrumentation and Procedures in Reading) at the University of Minnesota. A total of eleven measures are included in

earlyReading that expand beyond those included DIBELS, AIMSweb, and easyCBM. FAST earlyReading includes measures of LNF, LSF, PSF, NWF, sight word fluency (high frequency word reading), and a rendition of FSF that more closely resembled the original DIBELS ISF with pictures called Onset Sound Fluency (OSF). In addition, earlyReading includes measures of concepts of print and alternate measures related to the alphabetic principle, phonemic awareness and fluency (rhyming, word blending, decodable world fluency and sentence reading).

Specific Indicators of Early Reading

When interpreting estimates of reliability and validity for early reading indicators, it is important to keep in mind several things: 1) similarity of measures when generalizing (i.e., DIBELS FSF versus earlyReading Onset Sounds), 2) the timing at which the indicator was administered (i.e., beginning of kindergarten, middle of first grade) , 3) the criterion measure used to establish estimates (i.e. measure of broad reading achievement versus measure of phonological awareness), 4) the timing at which the criterion measure was administered (i.e., 6 months after indicator, 12 months after indicator), and 5) the established standards of reliability and validity for the specific purpose (i.e., high-stakes versus low-stakes decisions). Unlike reliability which has established standards for correlation coefficients equal to .70 for low-stakes decisions and .90 for high stakes decisions (Kelly, 1927), the gold standard for validity is less clear. In accord with APA standards, a rationale should be presented for each recommended interpretation (AERA, APA, & NCME, 1999). Consistent with standards used by DIBELS Next (Good, et al., 2011) and the National Center for Response to Intervention (NCRtI, 2012), coefficients

of at least .70 seem reasonable to use as standard of adequate validity for use with low-stakes decisions.

Concepts about Print (CAP). Rooted in the work of Marie Clay the original CAP measure evaluated children's concepts about book orientation, about whether print or pictures carry the text message, about directionality of lines of print, page sequences and directionality of words, about the relationship between written and oral language and about words, letters, capitals, space and punctuation (Clay, 1979a; Goodman, 1981). Clay's CAP included 24 items and used one of two 20-page booklets, Sand (Clay, 1972) and Stones (Clay, 1979b), to observe and evaluate these concepts. The measure is recommended for students at the beginning of the first year of reading instruction to help teachers plan appropriate reading experiences for children. Although Clay writes that it was not intended as a measure of reading readiness or designed to predict reading progress (Clay, 1989), researchers have used the CAP measure, or aspects of the concepts of print task, to distinguish groups of readers and develop models of word reading acquisition (Reutzel, Fawson, Young, Morrison, & Wilcox, 2003; Johns, 1980).

Onset Sound. The onset sound task is a measure of phonemic awareness. An onset is the consonant sound that precedes the word, where the rime is the vowel and any consonant sounds that come after it (Adams, 1994). For example, given the word "cat," /c/ is the onset and /at/ the rime. Given the word "shake," /sh/ is the onset and /ake/ the rime. A student's ability to distinguish and identify the onset from the rime has gained support in the literature as an indicator of both current and future reading achievement. Across six studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .20 to .71, with a median of .51, and from .14 to .60, with

a median of .39 for concurrent and predictive validity, respectively (Burke, Hagan-Burke, Kwok, & Parker, 2009; Cummings, Kaminski, Good, & O'Neil, 2010; Elliott, Lee, & Tollefson, 2001; Hintze, Ryan, & Stoner, 2003; E. S. Johnson, J. R. Jenkins, Y. Petscher, & H. Catts, 2009; Nelson, 2008).

Letter Naming. Research supporting the association between student knowledge of letter names and reading achievement is prevalent for use in both kindergarten and first grade. Across ten studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .32 to .84, with a median of .55, and from .37 to .74, with a median of .62 for concurrent and predictive validity, respectively (Burke, Crowder, Hagan-Burke, & Zou, 2009; Burke, Hagan-Burke, et al., 2009; Elliott, et al., 2001; Evans, Bell, Shaw, Moretti, & Page, 2006; Hintze, et al., 2003; Nelson, 2008; Ritchey, 2008; Rouse & Fantuzzo, 2006; D. Speece, Mills, Ritchey, & Hillman, 2003; Stage, Sheppard, Davidson, & Browning, 2001). In first grade, across seven studies, estimates ranged from .08 to .63, with a median of .47, and from .22 to .71, with a median of .40 for concurrent and predictive validity, respectively (Chard et al., 2008; Daly, Wright, Kelly, & Martens, 1997; Goffreda, Diperna, & Pedersen, 2009; Hagan-Burke, Burke, & Crowder, 2006; Landerl & Wimmer, 2008; Riedel, 2007; Schilling, Carlisle, Scott, & Zeng, 2007).

Letter Sound. The letter sound (LS) task is a measure of alphabetic principle. Letter sound knowledge has received much less attention in the literature than Letter Naming. As previously stated, a LSF (which includes a timing component) measure is included in AIMSweb, easyCBM, and FAST earlyReading, but the concept of measuring letter sound knowledge certainly did not originate just in the last 20 years (Gates, 1939).

Typically letter sound measures contain all 26 letters, with some measures displaying a total of 52 letters (26 upper-case, 26 lower-case). Still, other LS measures only include a subset of letters. Overall, LSF was judged to approach standards of reliability and validity for use as an indicator, and only slightly underperformed LNF. Across six studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .31 to .72, with a median of .58, and from .42 to .77, with a median of .67 for concurrent and predictive validity, respectively (Chafouleas, Lewandowski, Smith, Blachman, & 1997, 1997; Daly, et al., 1997; Elliott, et al., 2001; Evans, et al., 2006; Stage, et al., 2001; Walton, 1995).

Rhyming. Although rhymes are related to rimes, they are distinct. Rhyming is a comparison of rime units of two or more words. For example, "light" and "kite" share the same rime, where "missed" and "massed" do not despite word similarity. Across five studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .25 to .66, with a median of .42, and from .29 to .62, with a median of .45 for concurrent and predictive validity, respectively (Blaiklock, 2004; Chafouleas, et al., 1997; Cronin & Carver, 1998; O'Connor & Jenkins, 1999; Walton, 1995).

Word Blending. Word blending requires students to provide a word when given the phonemes orally. Based on estimates, word blending was also found to be associated with phonological awareness. Compared to rhyming, blending was determined to be a more difficult task; however, it was not the most difficult, with word segmenting and manipulation typically mastered later than blending (Chafouleas, et al., 1997). Across three studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .27 to .79, with a median of .52, and from .22 to .57, with a median

of .43 for concurrent and predictive validity, respectively (Chafouleas, et al., 1997; J. Chall, Roswell, & Blumenthal, 1963; O'Connor & Jenkins, 1999).

Word Segmenting. Word segmenting is another measure that has received a lot of attention in the literature. Where some measures, such as letter sound knowledge, and initial sound fluency are often examined only in regard to kindergarten, the examination of word segmenting extends from kindergarten into grade one. The word segmenting task is largely based off the work of Yopp (1988), where after a word is orally presented the student is asked to say the individual phonemes of the word. Across ten studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .09 to .73, with a median of .45, and from .27 to .75, with a median of .49 for concurrent and predictive validity, respectively (Burke, Crowder, et al., 2009; Burke, Hagan-Burke, et al., 2009; Chafouleas, et al., 1997; Elliott, et al., 2001; Hintze, et al., 2003; E.S. Johnson, et al., 2009; Nelson, 2008; O'Connor & Jenkins, 1999; Rouse & Fantuzzo, 2006; Spector, 1992). In first grade, across seven studies, estimates ranged from .08 to .40, with a median of .31, and from .08 to .43, with a median of .26 for concurrent and predictive validity, respectively (Burke & Hagan-Burke, 2007; Chard, et al., 2008; Goffreda, et al., 2009; Hagan-Burke, et al., 2006; E.S. Johnson, et al., 2009; Riedel, 2007; Schilling, et al., 2007).

Word Identification. For younger students who could not read connected text, reading word lists emerged as an alternative indicator of reading achievement. Using the word identification task, the difficulty of words presented can be more tightly controlled than the difficulty level of reading passages. Word lists can be created from a number of different word types including word structure (i.e., CVC or CVCe words) or frequency in

text (i.e. Dolch and Fry sight words). Although word lists can be characterized by decodable or sight words, the psychometric evidence available in the literature is focused on fluency with sight word lists. Decodable words lists more closely measure decoding skills including the use of letter sound correspondence and blending to produce a word (Carnine, et al., 2004), where sight word lists more closely measure word recognition skills including use of frequent spelling patterns to automatically translate a word to meaning (Adams, 1994). Across four studies evaluating the concurrent and predictive validity of sight words for kindergarten students, estimates ranged from .40 to .94, with a median of .72, and from .42 to .80, with a median of .69 for concurrent and predictive validity, respectively (Compton et al., 2010; Compton, et al., 2006; L. S. Fuchs, et al., 2004).

Nonsense Word Fluency. Nonsense Word Fluency (NWF) is a measure of pseudo-word reading. It was designed to isolate the ability to decode words from sight word reading. That is, with real words, it is unclear which skills students are using to read a word, namely decoding or sight word strategies. Similar to PSF and LNF, an extensive literature exists on the technical adequacy of NWF for use in both kindergarten and first grade. Across seven studies evaluating the concurrent and predictive validity for kindergarten students, estimates ranged from .24 to .91, with a median of .56, and from .29 to .77, with a median of .59 for concurrent and predictive validity, respectively (Burke, Hagan-Burke, et al., 2009; Fien et al., 2008; E.S. Johnson, et al., 2009; Nelson, 2008; Ritchey, 2008; Rouse & Fantuzzo, 2006; D. Speece, et al., 2003). In first grade, across fifteen studies, estimates ranged from .23 to .80, with a median of .66, and from .12to .86, with a median of .58 for concurrent and predictive validity, respectively

(Burke, Crowder, et al., 2009; Burke & Hagan-Burke, 2007; Chard, et al., 2008; Cummings, Dewey, Latimer, & Good, 2011; Fien, et al., 2008; Fien et al., 2010; L. S. Fuchs, et al., 2004; Goffreda, et al., 2009; Hagan-Burke, et al., 2006; Harn, Stoolmiller, & Chard, 2008; E.S. Johnson, et al., 2009; Riedel, 2007; Schilling, et al., 2007; D. Speece, et al., 2003; Vanderwood, Linklater, & Healy, 2008).

Composites. In an effort to increase the predictive validity estimates of pre-reading assessments, DIBELS Next developed a composite score with several early literacy measures for kindergarten and first grade (Good, et al., 2011). Using a combination of multiple DIBELS Next measures at the end of the first grade year, predictive-validity increased to .77, compared with a range of .40 to .75 for individual measures. That composite score is a weighted combination of student performances on NWF Correct Letter Sounds, NWF Whole Words Read, DIBELS Oral Reading Fluency (DORF) Words Correct, DORF Accuracy, and Retell. It is noted however, that estimates for composite scores are not always as robust as the .77 estimate found at the end first grade. At the beginning of the kindergarten year, for example, criterion-related validity estimates range from .52 for First Sound Fluency to .39 for LNF. The composite score estimate for the same grade level and time period is only .50. Similarly, for the beginning of the first grade year, individual estimates were equal to .54, .33, .43, and .39 for LNF, PSF, NWF Correct Letter Sounds, and NWF Whole Words Read, respectively, with a predictive validity estimate of .55 for the composite score. Given the variability in concurrent and predictive validity estimates of specific indicators, continued examination of composite scores is likely to strengthen estimates and functionality of these indicators.

Conclusion

Over the last century, theoretical perspectives on reading development have guided assessment practices related to reading. In the first half of the twentieth century when the study of reading was dominated by developmental perspectives (i.e., Gates and Dean), assessments often focused on reading achievement related to mental age. That is, students were assessed for their "readiness" to benefit from reading instruction. This is in contrast to an early intervention perspective of RTI which supports *more* instruction *earlier* for students judged to be at-risk of reading difficulties. With an early intervention perspective, these students can be identified using universal screening with an array of the early reading indicators described above.

Indicators of early reading, promote context specific evaluation as described by Reschly and Ysseldyke (2002) and Cronbach (1975). In the case of early reading, context specific evaluation refers to direct measurement of reading subskills taught in each unit of instruction (i.e., the lower order units associated with reading such as the alphabetic principle, phonemic awareness, and fluency). Reading assessments are no longer focused on "readiness," but instead on identifying students with skill deficiencies. Once students with specific skill deficiencies are identified, those skills can be directly taught and reassessed frequently using linked assessments. Universal screening using these indicators provides the basis for not only identifying students at the individual level who are in need of targeted services, but also to provide information about the functioning of the core curriculum (Ikeda, Neessen, & Witt, 2008; Shapiro, 2008)

Best practices in school psychology recommends that all universal screening tools be examined for efficiency and technical adequacy of the measure. Specifically, all

measures should be judged on the ability to identify potential problems, answer questions about efficacy of the core program, and be disaggregated and used by teachers (Ikeda, et al., 2008). As stated, GOMs, such as CBM-R, may not be appropriate for younger students who have not yet, or have only started to acquire pre-reading and beginning reading skills. In contrast, SSMM enable educators to assess more specific skills related to reading, such as phonemic awareness, phoneme-grapheme correspondence and word identification. Application of the repeatable, simple and efficient procedures used with CBM for indicators of early reading create viable options for universal screening for kindergarten and early first grade students, that are consistent with effective instructional practices and reading theory. As identified in the review of early reading indicators, the timing at which these indicators are administered (i.e., beginning, middle or end of kindergarten and first grade) and the criterion measure used along with the timing at which it is administered, all have implications for the validity and utility of early reading indicators. Given these consideration, the use of composite scores has potential to further enhance the reliability, validity, diagnostic accuracy, and instructional use of these tools to improve student learning outcomes.

Chapter 3

A series of initiatives that focus educators on the early identification of students who are at risk of reading difficulties have sparked increased research on best practices in universal screening for early readers ("IDEIA," 2004; National Reading Panel, 2000; "No Child Left Behind Act of 2001," 2001). While early reading assessments have been in use since at least the early twentieth century (Gates, 1926), improved tools and systematic implementation have advanced the identification process for all students. Still, the evaluation of assessment tools for screening purposes is relatively new. Glover and Albers (2007) identified several key features to evaluate screening assessments to include the appropriateness, technical adequacy, and usability of the screening tool. Best practices in school psychology reiterate these features and require that all universal screening tools be examined for efficiency and technical adequacy (Ikeda, et al., 2008). Kane (2006) further describe an argument-based approach to validity which includes a need for an explicit statement of the proposed interpretation, extended analysis in validation, and consideration of alternate interpretations of use.

Practical concerns dictate that screening measures be a cost effective sampling of reading skills which can be easily administered, scored and interpreted to inform instruction. Universal screening is typically administered up-to-three times per year to all students, but is sometimes only administered to a subset of students at a particular grade or classroom (Ikeda, et al., 2008). For elementary grades two through six, curriculum-based measurement of oral reading (CBM-R) is often used as a universal screener for grade-level reading proficiency. CBM-R requires students to read from a grade level passage for one minute while the number of words read correctly is recorded. Strong

evidence exists in the literature to support the use of CBM-R (L. S. Fuchs, et al., 1988; Kranzler, et al., 1998; Markell & Deno, 1997; Reschly, Busch, Betts, Deno, & Long, 2009; Wayman, Wallace, Wiley, Ticha, & Espin, 2007); however, support for CBM-R as a universal screener for students not yet reading connected text is less robust (L. S. Fuchs, et al., 2004; National Research Council, 1998). This difference is likely attributed to students' acquired level of reading skill.

Early Reading Skills

Before students start reading connected text, a number of pre-reading skills develop. A number of reading theorists point to acquisition of phonemic awareness, the alphabetic principle and/or decoding as key components in the development of reading (Chall, 1983; Ehri, 1999, 2005; Gates, 1949; Gough & Tunmer, 1986; W. A. Hoover & Gough, 1990; LaBerge & Samuels, 1974). In their review of effective instructional practices, the National Reading Panel (NRP, 2000) reiterated the importance of these early reading skills. The NRP highlighted acquisition of phonemic awareness and alphabetics in addition to fluency with connected text, vocabulary, and comprehension as important to becoming a good reader. Although vocabulary is implicated in the development of early reading, the validity of current vocabulary assessment practices with kindergarten and first grade students is questioned (Pearson, Haertel, & Kamil, 2007). As Pearson et al. (2007) suggested, research is still needed to determine whether vocabulary assessment requires a single broad based assessment, or more targeted assessments for different types of vocabulary.

Given the required advancement necessary in the conceptualization and measurement of vocabulary including explicit links to instruction, phonemic awareness

and the alphabetic principle are most often the targets of early reading indicators to date (Alonzo & Tindal, 2004; Good & Kaminski, 2002b; M. M. Shinn & Shinn, 2002). Outside of phonemic awareness and the alphabetic principle, assessments measuring concepts of print and decoding also show promise as indicators of early reading (Burke, Hagan-Burke, et al., 2009; Fien, et al., 2008; Hagan-Burke, et al., 2006; Johns, 1980; Lomax & McGee, 1987; Nelson, 2008; Speece, et al., 2003). Assessments of phonemic awareness, the alphabetic principle, concepts of print, and decoding align with curriculum and instruction, and adhere to the characteristics of universal screening measures which include ease of administration, scoring and interpretation, and are useful in identifying which students may need additional support. Concepts of print, phonemic awareness, alphabetic principle, and decoding are defined below.

Concepts about Print (CAP). Concepts of print is students' knowledge and understanding of print conventions. These include book orientation, whether print or pictures carry the text message, directionality of lines of print, page sequences and directionality of words, the relationship between written and oral language and of words, letters, capitals, space and punctuation (Clay, 1979a; Goodman, 1981). Rooted in the work of Marie Clay, the measure is recommended for students at the beginning of the first year of reading instruction to help teachers plan appropriate reading experiences for children. Although, Clay writes that it was not intended as a measure of reading readiness or designed to predict reading progress (Clay, 1989), researchers have used the CAP measure, or aspects of the concepts of print task, to distinguish groups of readers and develop models of word reading acquisition.

Phonemic Awareness. Phonemic awareness is the ability to recognize and manipulate phonemes, or individual sounds in spoken words. Phonemes are the smallest unit of sound in spoken language. For example, the word "cat" consists of three phonemes /c/, /a/, and /t/. Phonemic awareness assessment includes a variety of activities related to the ability to identify and manipulate phonemes in a spoken word. These assessments include less difficult tasks from identification of the first phoneme and blending to more difficult tasks such as segmenting whole words to deleting and adding phonemes.

Alphabetic Principle. The alphabetic principle is the relationship between the letters of a written language and the individual sounds (Adler, 2001; Moats, 2000). Acquisition of the alphabetic principle requires the ability to identify letters and sounds of the alphabet, along with their correspondence (National Research Council, 1998). Letter sound fluency measures are the most common type of measure to examine alphabetic principle (Chafouleas, et al., 1997; Daly, et al., 1997; Evans, et al., 2006; Walton, 1995)

Decoding. Decoding and word identification require students to read words in isolation. These assessments generally use lists of words, and require students to read for one minute. Word recognition skills include use of spelling patterns that occur frequently to automatically translate a word to meaning (Adams, 1994). In contrast, decoding skills include use of letter sound correspondence and blending to produce a word (Carnine, Silbert, Kame'enui, & Tarver, 2004). Decodable word lists consist of words with specific spelling patterns such as consonant-vowel-consonant (CVC) and allow students to sound out words using alphabetic principle knowledge (i.e., car, nut etc.). Decodable word lists

are often exclusive of words found in high frequency or sight word assessments. An alternative to the use of decodable words is use of nonsense word lists. Nonsense words allow direct assessment of student decoding skills using the same CVC spelling patterns, but use pretend words such as "nek" and "pof." In this assessment all words are unfamiliar to students so they must be decoded (instead of read by "sight"). Use of nonsense words is the most commonly researched, with fewer studies focused on sight or decodable words (Burke, Hagan-Burke, et al., 2009; Chard, et al., 2008; Compton, et al., 2006; Daly, et al., 1997; Fien, et al., 2008; L. S. Fuchs, et al., 2004; Hagan-Burke, et al., 2006; Johnson, Jenkins, & Petscher, 2010; Riedel, 2007; Ritchey, 2008; Rouse & Fantuzzo, 2006; Speece, et al., 2003).

Assessment Concerns

While useful in prediction and classifying students, concepts of print, phonemic awareness, alphabetic principle, and decoding are often mastered over relatively short periods of time (i.e. within one or two school years; Paris & Hoffman, 2004). This is especially true for skills that can be broken into smaller sub-skills such as phonological awareness. For example, in an examination of different phonological awareness assessments, Stanovich, Cunningham, and Cramer (1984) compared 10 assessments, ranging from rhyming to deleting initial consonants, and found assessments to range from easy to extremely difficult for kindergarten students based on distributions. This indicates that different phonological awareness measures may be more useful at different times within the kindergarten school year depending on difficulty level.

The short period of skill mastery is also reflected in the instructional priorities proposed by Simmons and Kame'euni (1999). Although, phonemic awareness is an

instructional priority throughout kindergarten, specific phonemic awareness skills may only be emphasized for several months during the school year. That is, while sound and word discrimination, rhyming, blending and segmenting are all instructional priorities of phonemic awareness at the beginning of the kindergarten, segmenting is the only phonemic awareness instructional priority by the end of kindergarten. Similarly, the instructional priorities of alphabetic principle which only includes letter-sound correspondence at the beginning of kindergarten shift to only include decoding and sight words by end of the kindergarten. A similar pattern of shifting priorities is observed in first grade with alphabetic principle and fluency with connected text.

This highlights the complexities involved in the assessment of early reading skills for kindergarten and first grade students. Skills are not taught in isolation, and nor is the emphasis on one type of skill consistent throughout a school year. As Morris, Bloodgood, Lomax and Perney (2003) suggest with their path analysis of reading skills across kindergarten and first grade, early reading skills likely develop in a symbiotic fashion where at some points reading achievement can be assessed using a single early reading skill, but at other time points the integration is so strong that assessment of multiple skills is recommended. Not surprisingly, given the number of skills associated with early reading, the National Research Council (1998) recommends using a combination of assessments for students who are beginning readers.

In lieu of using multiple assessments that yield multiple scores for teachers to interpret with little guidance, composite scores provide an aggregation of information to allow teachers to make systematic decisions about students (and instruction) based on multiple sources of data. Composite scores provide teachers with a single score to

summarize the combination of assessments (i.e., phonemic awareness, alphabetic principle, concepts of print, and decoding) that might be used during a single screening period. There are different ways to create composite scores including unit weighting and regression weighting. Although research on the use of composite scores is available across multiple disciplines, research for composites of reading and especially early reading is scarce (Bobko, Roth, & Buster, 2007). The DIBELS Management Group (DMG) use composite scores to provide the best estimate of reading proficiency at a given screening point (Good, et al., 2011); however DIBELS recognizes that because assessments used in the composite vary by time point, growth interpretations cannot be made across the school year. Rather composite scores are useful when rank ordering students, where a student's relative position can be compared across time (Bobko, et al., 2007). Research is needed to advance the utility of composite scores for reading in kindergarten and first grade.

Purpose

Universal screening measures that are aligned with curriculum and instruction in reading are necessary to evaluate the effectiveness of both the core curriculum and individual students in need of additional support. Despite extensive research on indicators of early reading, large scale studies examining the use of indicators for screening is less common. There is a need to examine multiple indicators in relation to other indicators at different time points throughout kindergarten and first grade.

The purpose of the current study was to examine a battery of early reading assessments for use as universal screeners in kindergarten and first grade. Specifically, twelve measures of early reading were developed. Each measure was quick and efficient

to administer, score and interpret, and developed to identify students in need of additional support in reading. The primary research questions are as follows:

1. To what extent does each earlyReading measure for fall, winter and spring have concurrent and/or predictive validity with a test of broad reading achievement given at the end of the school year in kindergarten and first grade?
2. To what extent can a composite of earlyReading measures be combined into valid composite scores to increase the concurrent and/or predictive validity for fall, winter and spring, respectively?

Methods

Participants

Kindergarten (N=233) and first grade students (N=180) from two school districts and six schools participated in the study. Two to three kindergarten classrooms (either half-day or full-day), and two to three first grade classrooms participated at each school. In District 1, the majority of students within the school district were White (53%), with the remaining students identified as African American (26%), Hispanic (11%), Asian (8%), or other (2%). Forty to fifty percent of students at each school received free and reduced lunch. On average, the majority of students across schools in District 2 were White (78%), with the remaining students identified as either African American (19%), or other (3%). Forty to fifty percent of students at each school received free and reduced lunch.

Measures

FAST earlyReading measures were developed as an extension of the Formative Assessment Instrumentation and Procedures for Reading (FAIP-R) project at the

University of Minnesota (Christ & Ardoin, 2009). earlyReading includes 12 subtests for pre-reading and early reading skills. All subtests use standardized procedures including prescribed directions that often include a practice section with standardized response sets and timed administration.

Concepts of Print. Concepts of Print was an untimed task where the student was presented with 12 items that measure knowledge of principles related to print conventions. Students were asked to identify printed numbers, letters, shapes and sentences. Students were then asked to distinguish between words of different length that contain the same root word. For example, when presented with the words "Roll" and "Rollercoaster," the student would be asked to point to the word "Roll." The measure of performance was number correct out of 12. Concurrent validity was equal to.60 with the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) standardized total score for kindergarten in the fall. Test-retest reliability was .42 for kindergarten in the fall (N=39).

Onset Sounds. Onset Sounds was an untimed task where the student was presented with a set of four pictures and asked to point to the picture with the same onset, or beginning sound, as a word provided by the examiner. A total of 16 items were included with distinct beginning sounds across the items. A fifth set of pictures was used for practice items. After the practice and training questions, students were asked to say or point to the picture with the same beginning sound as the prompted word. For example, the examiner would say while pointing to the four pictures, "This is a fish, train, zebra, and balloon. Which one begins with /z/?" The student would then receive credit for correctly pointing to or saying "zebra." The measure of performance was the number

correct out of 16. Concurrent validity was reported to equal .62 with the GRADE standardized total score for kindergarten in the fall. Test-retest reliability was .79 for kindergarten in the fall.

Letter Naming. Letter Naming was a partially timed task where the student was presented with a list of all 26 letters of the alphabet. Each letter was presented twice in both upper-and lower case. Students were asked to name each letter as quickly as possible without making mistakes. Although performance was timed for one minute students were encouraged to continue to name all the letters to provide a complete inventory of the student's letter naming accuracy. Although the measure of performance can be calculated by the total letters sounds read under timed and untimed conditions, in addition to the percent correct in one minute, the total number correct in one minute (rate score) was used as the measure of performance in the current study. Concurrent validity with the GRADE standardized total score for kindergarten in the fall was .41. Test-retest reliability was .91 for kindergarten in the fall.

Letter Sounds. Letter Sounds was very similar to Letter Names with the exception that students were asked to provide the sound of each letter instead of the name. Letters with dual sounds such as the vowels and "c" and "g" were presented at the bottom allowing for solicitation of both sounds. Dual sounds were administered during untimed conditions. Again, although the measure of performance could calculated by the total letters sounds read under timed and untimed conditions, in addition to the percent correct in one minute, the total number correct in one minute (rate score) was used as the measure of performance in the current study. Concurrent validity with the GRADE in the

fall was .53 for kindergarten students. Test-retest reliability was .75 for kindergarten in the fall.

Rhyming. The format was very similar to Onset Sounds, where the student was presented with five sets of four pictures. The first set was for training and practice, with the last four sets used for testing. For each set of pictures, three questions ask the student to say or point to the picture that rhymes with a given word. The fourth question asks the student to produce a rhyming word with a given picture. The measure of performance was the number correct out of 16. Concurrent validity with the GRADE in the fall was .58 for kindergarten students. Test-retest reliability was .74 for kindergarten in the fall.

Word Blending. Word Blending was an untimed task where the student was presented with a string of two to three sounds and asked to blend sounds to produce a word. For example, if the sounds /t/ /i/ /n/ were provided with a one second pause in between each sound, the student would have to successfully blend the sounds and produce the word "tin" to receive points. The measure of performance was the number correct out of 10. Concurrent validity with the GRADE in the fall was .22 for first grade. Test-retest reliability was .71 for kindergarten and .91 in first grade in the fall.

Word Segmenting. Word Segmenting was similar to Word Blending, except the student was provided a whole word and asked to produce the individual sounds. For example, if the word "tin" was provided, the student would respond with /t/ /i/ /n/. The measure of performance was the number of sounds correct out of 32. Concurrent validity with the GRADE in the fall was .49 for first grade. Test-retest reliability was .84 in first grade in the fall.

Nonsense Words. Nonsense Words was a timed task where the student was presented with a list of 50 nonsense words and asked to read from the list for one minute while errors were recorded. All words had the structure of consonant-vowel-consonant or vowel-consonant. The measure of performance was the number of words read correctly in one minute. Concurrent validity estimates were not available for Nonsense Words. Test-retest reliability was .84 in first grade in the fall.

Decodable Words. Decodable Words was a timed task where the student was presented with a list of 50 decodable words and asked to read from the list for one minute while errors were recorded. All words had the structure of consonant-vowel-consonant. The measure of performance was the number of words read correctly in one minute. Concurrent validity with the GRADE in the fall was .22 for first grade. Test-retest reliability was .95 in first grade in the fall.

Sight Word 50 and 150. Sight Word was a timed task where the student was presented with a list of 50 or 150 sight words and asked to read from the list for one minute while errors were recorded. Sight Words 50 contained all 50 words on one page for kindergarten students, where Sight Words 150 contained 50 words on each of 3 pages for first grade students. The measure of performance was the number of words read correctly in one minute. Concurrent validity with the GRADE in the fall was .59 for first grade. Test-retest reliability was .97 in first grade in the fall.

Sentence Reading. Sentence Reading was a timed task where the student was presented with a series of sentences and paragraphs, and asked to read quickly without making mistakes. The number of words read correctly in one minute was recorded. One sentence was presented on a single page for the first three sentences. Four sentences were

presented on the fourth page, with whole paragraphs presented on the fifth and sixth pages. In this way, only students who are successful with the limited text presented on the first three pages encounter the additional sentences on pages four through six. All sentences originate from a primer level FAIP passage (Christ & Ardoin, 2009), and contained a related picture on each page. The measure of performance was the number of words read correctly in one minute. Test-retest reliability was .98 for first grade in the fall.

CBMReading. CBMReading is a version of CBM-R created by *FAST*. Students read aloud from a page of text while words read correct and errors are marked and recorded. The passages developed for the Grade 1 passages used in this study included 150-200 words overall in 2-5 paragraphs. Sentence length ranged from 3 to 7 words, with each paragraph containing 7 to 15 sentences. The number of words per sentence and sentences per paragraph were varied across the story to result in the appropriate total number of words. For screening purposes, students read from three passages for one minute each. The measure of performance was the median number of words read in one minute across the three passages. Concurrent validity was .89 with the Test of oral silent reading and comprehension (TOSREC), .97 with AIMSweb, and .78 with DIBELSNext for first grade students. Predictive validity was .91 with AIMSweb for first grade students after 12 weeks. Test-retest reliability was .90 and .82 for first grade from fall to winter, and fall to spring, respectively. Inter-rater reliability ranged from .83 to 1.00, with a median of .97.

Group Reading Assessment and Diagnostic Evaluation (GRADE). The GRADE was a diagnostic screening tool used to determine the reading skills children

have mastered (Williams, 2001). The GRADE has eleven levels for use with students ranging from pre-kindergarten to young adulthood. Level K was administered to kindergarten students and Level 1 was administered to 1st grade students. Both assessments were group administered at the class level. The GRADE was administered in April and May of the school year. Split-half reliability coefficients corrected by Spearman-Brown formula were reported to range from .91 to .99. Criterion related validity ranged from .76 to .90.

Level K. Level K was designed as an early reading assessment for kindergarten, early first grade, and transitional first-grade classrooms. It consists of eight required subtests and one optional subtest. The subtests include Sound Matching, Rhyming, Same and Different Words, Print Awareness, Letter Recognition, Phoneme Grapheme Correspondence, Listening Comprehension and Word Reading. Together the subtests measure phonological awareness, visual skills, early literacy skills, knowledge of print material, basic early reading skills that require both visual and auditory skills, understanding of spoken language, recognition of basic pre-primer and primer sight words and decoding simple, regular words. All nine subtests were administered to kindergarten students. The measure of performance was the overall standard score based on the raw score of all nine subtests combined.

Level 1. Level 1 was designed as an early reading assessment for kindergarten, early first grade, and transitional first-grade classrooms. It consists of five required subtests. The subtests include Word Reading, Word Meaning, Sentence Comprehension, Passage Comprehension, and Listening Comprehension. Together the subtests measure vocabulary, comprehension and oral language. All subtests were administered to 1st grade

students. The measure of performance was the overall standard score based on the raw score of all five subtests combined.

Implementation Procedures

The data used in the study were obtained as a part of a pilot project of the FAST suite of assessments. School districts volunteered to administer a predetermined schedule of FAST earlyReading and CBMReading measures (see Table 1) three times per year to all students K-5 in exchange for full use of all FAST assessments at no cost. Data collection spanned one year and involved a cohort of kindergarten and first grade students.

earlyReading administration. Five to seven measures were administered to students in kindergarten and first grade at each screening period (fall, winter and spring) in the 2012-2013 academic year. On average, the screening battery took 10 to 20 minutes in kindergarten, and 8 to 12 minutes in first grade. These measures were all administered by classroom teachers and trained assistants. All teachers in District 1 were trained by district staff who attended a "train the trainer" online session. Teachers in District 2 attended a two-hour in-person training on the earlyReading measures and were observed a percentage of the time for administration integrity by the lead teacher at each school site. Estimates of administration integrity were not available.

GRADE administration. The GRADE was used as the criterion measure and group administered at the end of the school year by at least one graduate student and one additional adult. Sessions were broken into three 30 minute sessions for kindergarten, and two 45 minute sessions for first grade.

Analytic Procedure

The earlyReading measures for each screening period were identified as the predictor variables. GRADE standard scores were used as the outcome variable. Distribution and residual plots were examined for each of the predictor variables at each screening period to review the five assumptions required for estimation and inference. Transformations were performed as necessary to meet the assumption of linearity. First, simple regression models were fit for each of the predictor variables. Data were then analyzed using linear multiple regression for each grade level across the fall, winter, and spring. Procedures similar to Speece et al. (2011) were used where all-subset regression identified the subset of the most efficient predictors. The top three models using each of one to five predictors were reported for each season. Results were limited to the top three models from each subset due to practical concerns. All models were first compared to the full model using a 2 percent difference in explained variation as the rule of thumb for identifying differences between models (Cohen, 1988).

Models were also compared against criteria used by the National Center for Response to Intervention (NCRtI), where correlations estimates equal to, or above .70 were considered to have "convincing evidence" for validity when used as a universal screener (NCRtI, 2012). Two methods were used to create composite scores and evaluated to include regression based weighting and unit weighting (Bobko, et al., 2007). Unit weighting refers to summing standardized scores (i.e. scores are converted to z-scores before applying equal weights). This is in contrast to summing raw scores. The assumption of dimensionality of the construct of reading in kindergarten and first grade

was evaluated using confirmatory factor analysis at each time point (fall, winter, and spring).

Results

Results for kindergarten are presented for fall, winter, and spring, followed by first grade. Within the present study, differences were not found between regression and unit based weighting; therefore, results were not presented for unit weighting. The assumption of uni-dimensionality was tested using confirmatory factor analysis. Uni-dimensionality was found in the fall and winter of kindergarten; however, bi-dimensionality was found in the spring of kindergarten through spring of first grade. There was a factor for word reading which included Nonsense Words, Sight Words and Decodable Words and another for phonemic awareness tasks which included Word Blending and Word Segmenting. In the spring of kindergarten, Letter Names and Letter Sounds were included in both factors.

Kindergarten.

The means, standard deviation and range for scores on the GRADE and each of the earlyReading measures by season for kindergarten are presented in Table 2. The predictors included in the kindergarten analysis varied by season (see Table 1). An examination of the distribution plots for each earlyReading measure suggested floor and ceiling effects were present for some measures throughout kindergarten; however, given the extensive schedule of administration, many of these effects were expected. All kindergarten earlyReading measures met the assumption of linearity based on examination of the studentized residual plots, with the exception of Nonsense Words in the spring, as the residual values were systematically under-predicted at the high end of

the distribution. After a log transformation was applied, Nonsense Words met the assumption of linearity. The log of Nonsense Words was used in place of Nonsense Words for all remaining analyses.

Similarly, all kindergarten earlyReading measures met the assumption of independence, as the GRADE standard scores were independent of each of the earlyReading measures. The assumption of homoscedasticity was met for all earlyReading measures as the error variance did not systematically increase or decrease across any of the distributions. Lastly, for all kindergarten earlyReading measures, less than 5% of the observations had residual variances that fell more than two standard deviations from the mean. The correlations between the outcome measure and all required predictor variables were in the range of .41 to .59 in the fall, .46 to .62 in the winter and .43 to .54 in the spring (see Table 3). Multicollinearity was evaluated using the variance inflation factor (VIF). A VIF greater than 10 was used as an indication of multicollinearity (J. Cohen, Cohen, West, & Aiken, 2003). VIF values ranged from 1.49 to 6.45 across the school year for kindergarten; therefore multicollinearity did not appear to excessively influence the results.

Simple regression models were fit for each predictor in the fall, winter, and spring to predict standard scores on the GRADE at the end of the school year for kindergarten students. For predictors that were required as part of the study at each time point, R^2 ranged from .17 to .34, .22 to .39, and .18 to .35 in the fall, winter, and spring, respectively. All predictors were statistically significant in the simple regression models. In the fall, Rhyming ($r^2 = .35$, $F(1, 212) = 112.1$, $p < .001$), Onset Sounds ($r^2 = .31$, $F(1, 212) = 94.51$, $p < .001$), and Concepts of Print ($r^2 = .25$, $F(1, 212) = 70.49$, $p < .001$)

were the top three predictors. Using a 2 percent difference as the rule of thumb for identifying differences between models (Cohen, 1988), Rhyming ranked as the number one predictor, Onset Sounds as the second, and Concepts of Print as the third. In the winter, Rhyming (r^2 = .39, F(1, 207) = 131.5, p < .001), Letter Sounds (r^2 = .33, F(1, 193) = 93.63, p < .001), and Word Segmenting (r^2 = .33, F(1, 211) = 102.4, p < .001) were the top three predictors. Again, Rhyming ranked first, with no differences found between Letter Sounds, and Word Segmenting in the winter. In the spring, Nonsense Words (r^2 = .35, F(1, 214) = 115.3, p < .001), Rhyming (r^2 = .29, F(1, 213) = 86.95, p < .001), and Decodable Words (r^2 = .27, F(1, 213) = 78.73, p < .001) were the top three predictors. Nonsense Words ranked as the number one predictor, Rhyming ranked second, and Decodable Words ranked third.

Using all-subset regression, models with the best three combinations of predictors were identified using two to five predictors each for fall, winter, and spring (see **Error! Reference source not found.**). In the fall, models with subsets of three to four predictors were found comparable to the full model which used all six predictors (Concepts of Print, Onset Sounds, Letter Names, Letter Sounds, Rhyming and Word Blending) based on explained variation. For example, the full model (R^2 = .47, F(6, 207) = 30.28, p < .001) only added 1.1 percent explained variance to the 46 percent variance accounted for when using the best subset of three predictors (Concepts of Print, Onset Sounds, and Rhyming; R^2 = .46, F(3, 210) = 58.82, p < .00).

In the winter, similar patterns were observed where models with three to four predictors were found comparable to the full model using all six predictors (Onset Sounds, Letter Names, Letter Sounds, Rhyming, Word Blending, and Word Segmenting;

R^2 = .56, F(6, 173) = 37.24, p < .001). The best fitting model using three (Letter Sounds, Rhyming and Word Segmenting; R^2 = .54, F(3, 189) = 74.38, p < .001) and four predictors (Onset Sounds, Letter Sounds, Rhyming and Word Segmenting; R^2 = .56, F(4, 175) = 56.26, p < .001), respectively, explained 2.1 and 0.1 percent less variance than the full model.

In the spring, again similar patterns were observed where models with three to four predictors were found comparable to the full model using all six predictors (Letter Names, Letter Sounds, Rhyming, Word Blending, Word Segmenting, Nonsense Words, Sight Words, and Decodable Words (R^2 = .51, F(8, 203) = 26.31, p < .001). The best fitting model using three (Rhyming, Word Blending, and Nonsense Words; R^2 = .48, F(3, 208) = 63.48, p < .001) and four predictors (Rhyming, Word Blending, Nonsense Words and Decodable Words; R^2 = .50, F(4, 207) = 51.07, p < .001), respectively, explained 3.0 and 1.2 percent less variance than the full model.

First Grade.

The means, standard deviation and range for scores on the GRADE and each of the earlyReading measures by season for first grade are presented in Table 4. The predictors included in the first grade analysis were fairly consistent across seasons with Word Blending, Word Segmenting, Decodable Words, Sight Words 150 and Nonsense Words administered each time. In contrast, Sentence Reading was administered in the fall, where CBMReading was administered in winter and spring (see Table 1). It is noted, however, that due to missing data, Nonsense Words was excluded from spring analysis. An examination of the distribution plots for each earlyReading measure suggested floor and ceiling effects for measures of phonemic awareness throughout first grade; however,

given the extensive schedule of administration, these effects were expected. All first grade earlyReading measures met the assumption of linearity based on examination of the studentized residual plots, with the exception of Sentence Reading and CBMReading in the fall and winter respectively, as the residual values were systematically under-predicted at the high end of the distribution. After a log transformation was applied, Sentence Reading and CBMReading met the assumption of linearity in the fall and winter respectively. The log of Sentence Reading and CBMReading in the fall and winter were used in place of Sentence Reading and CBMReading for all remaining analyses.

Similarly, all first grade earlyReading measures met the assumption of independence as the GRADE standard scores were independent of each of the earlyReading measures. The assumption of homoscedasticity was met for all earlyReading measures as the error variance did not systematically increase or decrease across any of the distributions. Lastly, for all first grade earlyReading measures, less than 5% of the observations had residual variances that fell more than two standard deviations from the mean. The correlations between the outcome measure and all required predictor variables ranged from .32 to .66 in the fall, .41 to .76 in the winter and .27 to .82 in the spring (see Table 5). Multicollinearity was again evaluated using the VIF. VIF values ranged from 1.39 to 6.48 across the school year for first grade; therefore multicollinearity did not appear to excessively influence the results.

Simple regression models were fit for each predictor in the fall, winter, and spring to predict standard scores on the GRADE at the end of the school year for first grade students. For predictors that were required as part of the study at each time point, R^2 ranged from .11 to .52, .17 to .67, and .07 to .68 in the fall, winter, and spring,

respectively. All predictors were statistically significant in the simple regression models. In the fall, Sentence Reading (R^2 = .52, F(1, 171) = 187, p < .001), Sight Words (R^2 = .44, F(1, 171) = 133.2, p < .001), and Nonsense Words (R^2 = .36, F(1,171) = 94.5, p < .001) were the top three predictors. Sentence Reading ranked as the number one predictor, Sight Words as the second, and Nonsense Words as the third best predictor based on R^2. In the winter and spring, CBMReading (R^2 = .67, F(1, 175) = 360.1, p < .001 and R^2 = .68, F(1, 180) = 378.9, p < .001, respectively), Sight Words (R^2 = .54, F(1, 158) = 188.7, p < .001 and R^2 = .43, F(1,165) = 123.3, p < .001, respectively), and Decodable Words (R^2 = .52, F(1, 159) = 175.4, p < .001 and R^2 = .43, F(1, 178) = 132.3, p < .001, respectively), were the top three predictors. CBMReading ranked first in both winter and spring. In the winter, differences were found between Sight Words and Decodable Words with Sight Words out performing Decodable Words; however, by spring, these differences were negligible.

Using all-subset regression, models with the best three combinations of predictors were identified using two to five predictors each for fall, winter, and spring (see Figure 2). In the fall, models with two to three predictors were found comparable to the full model using all six predictors (Word Blending, Word Segmenting, Nonsense Words, Sight Words, Decodable Words and Sentence Reading; R^2 = .59, F(6, 166) = 39.98, p < .001) based on explained variation. The best fitting model using two predictors (Word Blending and Sentence Reading; R^2 = .58, F(2, 170) = 115.2, p < .001), explained 1.6 percent less variation than the full model. By winter, models containing CBMReading and one additional predictor were consistently comparable to the full model. The full model for winter included Word Blending, Word Segmenting, Nonsense Words, Sight

Words, Decodable Words and CBMReading. When including only Word Segmenting and CBMReading in the model, $R^2 = .69$, $F(2, 158) = 178.8$, $p < .001$ compared to $R^2 = .71$, $F(6, 148) = 61.08$, $p < .001$ for the full model. Compared to the full model, CBMReading and Word Blending together explained 2.2 percent less variation. By spring, CBMReading alone was comparable to the full model. The full model for spring included Word Blending, Word Segmenting, Sight Words, Decodable Words and CBMReading. Again, Nonsense Words was excluded from spring analysis due to low administration rates. When including CBMReading alone in the model $r^2 = .68$, $F(1, 180) = 378.9$, $p < .001$ compared to $R^2 = .70$, $F(5, 160) = 73.57$, $p < .001$ for the full model. Compared to the full model, the model including only CBMReading explained 1.9 percent less variation.

Discussion

The purpose of this study was to identify a screening battery at each time point for the fall, winter, and spring of kindergarten and first grade. The concurrent and predictive validity of each predictor, and varying composites were examined. In kindergarten screening batteries of three to four measures were found comparable across time points to the full screening battery consisting of six to seven measures. In first grade, screening batteries of two to three measures in the fall, and one to two measures in the winter and spring were found comparable to the full screening battery consisting of five to six measures.

Specifically in the fall of kindergarten, Rhyming, Onset Sounds and Concepts of Print were the top three single predictors of GRADE performance at the end of the school year; however, as single predictors all three continued to fall short of the NCRtI criteria

of $r \geq .70$ ($R^2 = .49$) for universal screeners of early reading. Similar findings were observed in the winter and spring of kindergarten as the highest r^2 for single predictors equaled .39 for Rhyming in the winter and .35 for Nonsense Words in the spring. The top three predictors were Rhyming, Letter Sounds, and Word Segmenting in the winter and Nonsense Words, Rhyming and Decodable Words in the spring.

When using multiple predictors based on weighted composite formulas, values of R^2approached the NCRtI criteria in the fall, and exceeded the NCRtI criteria in the winter and spring when the full battery of measures was used. Reduced composites consisting of fewer measures across the kindergarten year were found comparable to the full composite. Although it was recognized that several different composites at each time point were comparable, the models in Table 6 were identified at each time point that upheld or approached the standards of criterion validity and were also aligned to different domains of early reading (ie. concepts of print, phonemic awareness, alphabetic principle and decoding). These include composites of Concepts of Print, Onset Sounds, Letter Sounds and Rhyming (Model A) in the fall, Letter Sounds, Word Segmenting and Rhyming (Model B) in the winter, and Word Segmenting, Rhyming, Nonsense Words, and Decodable Words (Model C) in the spring.

Educators could justify the use of a different composite with comparable criterion validity. For example, Model D provides an alternative example of the kindergarten spring composite where Letter Sounds is used in place of Nonsense Words; however, given the composite of Letter Sounds, Word Segmenting, Rhyming, and Decodable Words it is noted that Letter Sounds was not a significant predictor, but likely a desirable

measure for instructional utility. As identified in Table 7, different composites may include or exclude certain domains depending on the measures included in the composite.

Overall in kindergarten, the composites chosen in this study vary between three and four measures and can be administered in approximately 8 to 12 minutes per student. The full screening battery consisted of the administration of at least six measures in the fall and winter, and seven measures in the spring requiring 10 to 20 minutes per student. Although total administration times for the full battery as observed in this study were lower than those observed in previous studies that used screening batteries (Catts, Fey, Zhang, & Tomblin, 2001; O'Connor & Jenkins, 1999), universal screening programs that requires more than 10 minutes per student is a costly practice which might not be sustainable. Efficient screeners with moderate to high concurrent and predictive validity are most desirable. Continued improvements of kindergarten early reading assessments are required.

In first grade, more promising results were observed. Across fall, winter and spring of the first grade year, Sentence Reading and CBMReading used as single predictors met the NCRtI, criteria of $r \geq .70$ ($R^2 = .49$). In addition to the reading of connected text, measures of word reading (Sight Words and Decodable Words) also met the criteria in the fall and winter, but not the spring. It is noted that although Sight Words and Decodable Words met standards, Sentence Reading/CBMReading were significantly better single predictors. When predictors were combined in a weighted composite, a marked increase in predictive validity was observed in the fall when a phonemic awareness measure was included with Sentence Reading. An increase in R^2 from .52 with Sentence Reading alone to .55 using Sentence Reading and Word Segmenting was

observed in the fall. In the winter and spring, smaller differences were observed between CBMReading, and CBMReading and Word Segmenting. That difference in R^2 equaled .02 in the winter and .01 in the spring, suggesting that as the first grade year progresses, the validity of CBMReading is strengthened as a very useful indicator of end of the year performance. Given that, use of Sentence Reading and Word Segmenting were selected as the composite for fall, with use of only CBMReading in the winter and spring for first grade (See Table 8). Similar to kindergarten, educators could choose different measures (i.e. Sentence Reading and Word Blending in the fall) and continue to meet standards of criterion validity.

Implications

Results were consistent with National Reading Council (1998) which recommended multiple indicators to assess beginning reading skills in young students; however, this was in contrast with findings of Morris et al. (2003), who suggested that reading developed in a symbiotic fashion where at some points reading achievement could be assessed using a single early reading skill, but at other time points, the integration is so strong that assessment of multiple skills is recommended. Results from the current study provide evidence of a systematic shift from multiple indicators of early reading to single indicators by the end of first grade. Likewise, results are generally consistent with instructional priorities identified by Simmons and Kame'enui (1999), in that there is a substantial utility for measures of phonemic awareness (i.e., Onset Sounds, Rhyming, Word Blending and Word Segmenting) and alphabetic principle (i.e., Letter Sounds) in early kindergarten, with shifting emphasis to reading connected text by end of

grade one. This consistency highlights the general alignment between the FAST earlyReading measures and curriculum and instruction.

In kindergarten use of single predictors continued to produce insufficient values of concurrent and predictive validity. Use of multiple predictors within a weighted composite show more promising results and improve upon estimates of concurrent and predictive validity of DIBELS composites (Good, et al., 2011), but continue to fall short of criteria for universal screening in the fall. Performance of the Concept of Print measure in the fall was encouraging, as it outperformed measures of Letter Names, Letter Sounds, Onset Sounds, and Word Blending. Although Concepts of Print is a newer measure within universal screening batteries for early reading, the FAST measure appears useful to identify students who are at-risk of reading difficulties prior to formal reading instruction. That is, as suggested by Catts et al. (2009), measures of onset sounds and letter naming fluency may be particularly sensitive to instruction. This is in contrast to the general concepts surrounding reading that students enter kindergarten with which may be a good indication of who will respond to instruction and who might not. This implies that Concepts of Print may overcome the limitation of other measures of early reading that require at least some formal reading instruction (Catts, et al., 2009). It is noted though, that the usefulness of this measure is one that likely diminishes quickly with instruction.

Based on the results from first grade, it is clear that Sentence Reading is a strong predictor in the fall of first grade. Similar to CBM-R, Sentence Reading measures the number of words read correct using connected text; however, in contrast to previous findings where CBM-R was a weak predictor in the beginning of first grade (Catts, et al.,

2009; Johnson, et al., 2010; Riedel, 2007; Wayman, et al., 2007), FAST Sentence Reading was a robust predictor in the fall of first grade. The presentation and format of the Sentence Reading measure were more in line with the types of connected text commonly used in first grade classrooms (Foorman, Francis, Davidson, Harm, & Griffin, 2004; Hoffman et al., 1994; Jenkins, Peyton, Sanders, & Vadasy, 2004; Juel & Roper-Schneider, 1985; Mesmer, 2005; Rhodes, 1981), where only one sentence is initially presented on a page with use of bigger font and a basic picture. As students with higher reading ability progress through the one minute measure, pages with multiple sentences are encountered. Likewise, the text difficulty level of FAST Sentence Reading was highly controlled with writing specifications tailored to the lowest level of reading ability (Pratt et al., 2011). In this way, the floor effects of screening batteries are limited compared to previous research, while maintaining a sufficient ceiling (Catts, et al., 2009). Using Sentence Reading and one additional measure such as Word Blending or Word Segmenting, estimates of predictive validity met the criteria for universal screening in the fall and exceed estimates found using DIBELSNext composites in the fall of first grade (Good, et al., 2011). In addition, administration time for a battery of two measures is reduced to approximately two minutes per student compared to the 8 to 12 minutes for the full battery.

By winter of first grade, the addition of a second measure, such as Sight or Decodable Words, to the universal screening battery for first grade children provides little increase to the predictive or concurrent validity. Consistent with findings of Riedel (2007), additional measures outside of FAST CBMReading in the spring did not explain additional variance. This highlights the usefulness of CBMReading to rank students in an

increasingly consistent manner across the first grade year when compared to a more lengthy diagnostic measure such as the GRADE used in this study. FAST composites again meet criteria for universal screening in the winter and spring and exceed estimates found from DIBELSNext composites (Good, et al., 2011). As the administration of three CBMReading passages are recommended in practice (M. R. Shinn, 2002, 2008), total administration time for universal screening in winter and spring of first grade is just over three minutes.

Limitations and Future Directions

While the current study highlights the value in using weighted composite formulas to reach minimal standards of concurrent and predictive validity in kindergarten and first grade, results should be interpreted in light of existing limitations. First, negative implications for comparing composites across time are associated with the break in uni-dimensionality observed in this study between the winter and spring of kindergarten. Although the validity of composite scores are not compromised when multiple components are measured within the composite, in this case phonemic awareness and word reading (Kane & Case, 2010), comparison of students' composite score to evaluate student growth across seasons becomes more difficult when those components change over time. Future research should focus on the dimensionality of reading within kindergarten and first grade to determine better methods to compare growth across season.

Additional limitations include the sample from which the criterion related estimates were obtained as generalization of results may be limited. All schools were from the Midwest, with classroom participation based on schools and teachers who

volunteered. The two districts involved this study were also trained under different training models. While both training models were considered of high quality, differences in administration could be a source of error. Likewise, the data analysis methods used were highly data-driven and therefore subject to sampling error. Future research should focus on the cross-validation of results.

Similarly, despite efforts to ensure adherence of the test administration schedule, some teachers failed to administer required measures at each time point, namely Nonsense Words. The test administration schedule also contributed to skewed distributions. That is, because "difficult" measures were administered earlier than usual and "easier" measures administered later than usual, the presence of floor and ceiling effects were observed for some measures. Likewise, the inclusion of additional tests at each time point could increase the concurrent and predictive validity of universal screening. For example, the administration of Sentence Reading in the spring of kindergarten might explain additional variance beyond the phonemic awareness and word reading measures used in this study. Other measures, such as vocabulary as suggested by Pearson et al. (2007) would likely increase validity estimates. Future research should continue to establish an appropriate set of measures for universal screening of early reading that are efficient, technically adequate and linked to instruction.

Of course, concurrent and predictive validity are only one aspect of the validity argument for use of composites within universal screening for early reading (Kane, 2006). Despite convincing evidence for concurrent and predictive validity based on correlational data, students are being misclassified as "at-risk" or "not at-risk" at higher rates than is acceptable within education (Catts, et al., 2009; Hintze, et al., 2003; Johnson,

et al., 2009; Nelson, 2008). Additional research is needed to establish validity including the diagnostic accuracy of composites.

The criterion used in this study also influences results. While it is the assumption that the standard is 100% accurate, use of a different measure and criterion performance level might produce different results. For example, as Jenkins, Hudson, and Johnson (2007) highlight, the outcome criterion used in similar studies varies drastically with use of the Woodcock Johnson Test of Achievement – Revised, Woodcock Reading Master Test, Stanford Achievement Test, 10th ed., CBM-R, or the double discrepancy commonly used in kindergarten and first grade research. Likewise, the criterion performance level used by NCRtI may be useful for comparison, but is somewhat arbitrary. Where standards of reliability are established in the literature (Cohen, 1988), standards of validity are less so. For example, where NCRtI states the performance level with no reference (NCRtI, 2012), others cite non-peer reviewed websites (Good, et al., 2011; 2002); both however describe correlation coefficients of .70 or above as "strong" or "convincing evidence."

Conclusion

Despite these limitations, this study highlights the need to continue the pursuit of appropriate, technically adequate, and usable screening batteries for identifying students at risk of reading difficulties in kindergarten and first grade students (Glover & Albers, 2007). In this study, quick and efficient screening batteries were established with adequate concurrent and predictive validity at most time points throughout kindergarten and first grade. It is clear that multiple measures are required within early reading screening, especially in kindergarten. Continued research is required to solidify the

combination of measures necessary for maximum predictive validity and diagnostic accuracy.

Chapter 4

The response to intervention (RTI) framework, using a multi-tiered system, supports early identification and intervention for students at risk of reading difficulties. The RTI framework relies on quality core instruction, universal screening, increasingly intense tiers of support, progress monitoring and data-based decision making. Although all are important, at the crux of RTI is universal screening (Jenkins, et al., 2007). Universal screening is used to evaluate the effectiveness of core instruction and identify students in need of intervention. Glover and Albers (2007) identified several key features to evaluate screening assessments to include the appropriateness, technical adequacy, and usability of the screening tool. Beyond correlational data which historically provided sole evidence of a screening measure's criterion validity, the degree to which a measure distinguishes between students at-risk and not at-risk for poor reading outcomes is a key component within test validation. Conceptions of validity have gradually shifted from the validation of the test itself to the validation of the proposed *interpretation of use* (Kane, 2006). That is, much like program evaluation, test validation requires a program to evaluate the "adequacy and appropriateness of inferences and actions based on the test scores." (Messick, 1989; p.13). Now known as an "argument-based approach to validity," an explicit statement of the proposed interpretation, extended analysis in validation, and consideration of alternate interpretations of use are required for validation.

Indicators of reading can be used for different purposes including universal screening, progress monitoring and diagnostic evaluation. With each purpose, the technical adequacy and appropriateness of inferences and actions change. In general, as highlighted by Fuchs, Fuchs, Hosp and Hamlett (2003), the adequacy and appropriateness

of curriculum based measurement (CBM) for use as a universal screener and to monitor individual student progress is more advanced than the adequacy and appropriateness of the same tool for other purposes. Curriculum based measurement of oral reading (CBM-R) is a common CBM that requires students to read from a grade level passage for one minute while errors and the number of words read correct are recorded. Correlational data and decision accuracy statistics contributed to the establishment of CBM-R as a reliable, valid, and efficient tool for universal screening in reading for grades two through six (Johnson, et al., 2010; Wayman, et al., 2007). Given the acquired level of reading skill by students in first grade and below, it is not surprising that evidence for the technical adequacy and appropriateness of CBM-R as a universal screener is less robust for younger students (L. S. Fuchs, et al., 2004; National Research Council, 1998). Alternatives to CBM-R are emerging for younger students. These include measures of concepts of print, phonemic awareness, alphabetic principle, and decoding (Alonzo & Tindal, 2004; Clay, 1989; Good, et al., 2011; Good & Kaminski, 2002b; M. M. Shinn & Shinn, 2002). The technical adequacy and appropriateness of inferences of these measures must be evaluated within the context of universal screening for kindergarten and first grade students.

Classification Accuracy

Within the context of universal screening for early literacy, decision accuracy refers to the degree a tool accurately identifies students who struggle with reading comprehension in the later grades. This includes students with low reading achievement and those identified with reading disabilities. Two correct classifications are possible. The first includes students who continue to struggle with later reading comprehension

and are identified as “at-risk” of reading difficulties. The second includes those who do not struggle with later reading comprehension and are identified as not “at-risk.” These are called true positives and true negatives, respectively. Incorrect classifications are also possible, known as false positives and false negatives. False positives occur when students are identified as “at-risk” of reading difficulties, but do not struggle with later reading comprehension. False negatives occur when students are identified as not “at-risk” of reading difficulties, but struggle with later reading comprehension. Decision accuracy can also be discussed in terms of area under the curve (AUC), sensitivity and specificity. AUC is a summary statistic of prediction accuracy and is not associated with a given cut point. This is in contrast to sensitivity and specificity which are generic to any number of cut-points. Sensitivity refers to the proportion of truly “at-risk” students who are identified as being “at-risk,” where specificity refers to the proportion of truly not “at-risk” students who are identified as being not “at-risk” (see Figure 3). Within early literacy research on universal screening, sensitivity and specificity are most commonly reported.

Due to the error associated with all tests, there is usually a trade-off associated with sensitivity and specificity. When cut points are adjusted to ensure all students “at-risk” are identified (i.e. high sensitivity), it generally follows that the number of false-positives also increases (i.e. lower specificity). Likewise, when cut points are adjusted to ensure high specificity, sensitivity is decreased. Improvements in screening tools are reflected by increases in both of these statistics. Where a desirable screening tool has high sensitivity and specificity, a poor screening tool is associated with poor sensitivity

and/or specificity. As no screening tool is perfect, educators and researchers must determine the acceptable levels for each.

Within educational research, sensitivity is often emphasized over specificity. For example, where Speece and Case (2001) explicitly state a bias for sensitivity when interpreting results, others choose cut points associated with high levels of sensitivity (i.e. .95 to 1.00) with little regard for specificity (Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; O'Connor & Jenkins, 1999). While lower sensitivity thresholds of .80 to .90 allow for more acceptable levels of specificity, Silberglitt and Hintze (2005) advocate for a more balanced approach to sensitivity and specificity. In the balanced approach, sensitivity and specificity are brought to .70 when possible (with preference given to sensitivity if not possible) and then continually increased in a step-wise procedure until a reasonable balance is achieved. When sensitivity can no longer be increased without decreases in specificity, the cut-point is selected. In this way, the overall classification accuracy, which accounts for both sensitivity and specificity, is often maximized.

Classification Accuracy of Early Reading Measures

Hintze, Ryan, and Stoner (2003) examined the diagnostic accuracy of several Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good, et al., 2011). The DIBELS assessment system included five pre-reading measures that are simple and efficient to administer including Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), Phonemic Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF). Hintze et al. examined levels of sensitivity and specificity for ISF, PSF and LNF when using the Comprehensive Test of Phonological Processing (CTOPP) as the criterion in kindergarten. High levels of sensitivity, but insufficient levels

of specificity were found for the recommended cut score by DIBELS. Using alternate cut points, the authors determined that DIBELS ISF and DIBELS LNF resulted in adequate levels of both sensitivity and specificity according to the .75 or higher standard (Swets, 1988). That study provided support for the use of DIBELS ISF and LNF for screening purposes, but not high-stakes diagnostic decisions. Nevertheless, consensus of cut points and screening practices have yet to emerge in the literature.

When using similar procedures and analysis, Nelson (2008) concluded that classification accuracy for DIBELS ISF, PSF, LNF, and NWF were still not sufficient for screening purposes in kindergarten. In that study the Woodcock-Johnson Tests of Achievement – Third Edition (WJ-III) and the second edition of TOPA (TOPA-2+) were used as criterions, with a more rigorous standard of .90 for sensitivity. A third study examined early versions of DIBELS LNF and NWF measures in kindergarten. Using oral reading fluency (ORF) as the criterion, Speece et al. (2003) determined that NWF and to a lesser extent LNF were valid measures for screening of early reading risk status beginning at the end of kindergarten. In that study specific standards for acceptable sensitivity and specificity were not reported.

There are multiple published recommendations for use of multiple early reading measures for screening in kindergarten and first grade. Multiple sources of information can improve AUC along with sensitivity and specificity estimates (Johnson, et al., 2010; Johnson, Jenkins, Petscher, & Catts, 2009). Johnson et al. (2010) describe two methods. The first approach uses a multivariate or combination of measures where students are identified as “at-risk” of reading difficulties using multiple cut points across several measures. For example, Johnson et al. (2009) used cut points for both ORF and the

Peabody Picture Vocabulary Test (PPVT) to determine risk status. First grade students scoring at or below 14 words read correct on ORF and below 109 on the PPVT at the beginning of first grade were classified as "at-risk." These cut points were determined because all students scoring at or below 14 words read correct on ORF, *and* equal to or above 109 on the PPVT performed above the 20th percentile on the Stanford Achievement Test (SAT) at the end of first grade; thereby reducing the number of false positives.

Using a similar approach known as classification tree analysis, Compton, Fuchs, Fuchs, and Bryant (2006) examined a screening battery in the fall of first grade. Classification tree analysis uses a series of decision rules governed by if-then logical conditions (Breiman, Friedman, Olshen, & Stone, 1984). The screening battery included sound matching, rapid digit naming, oral vocabulary and five weeks of progress monitoring with word identification fluency (WIF). The criterion was a composite of untimed word identification and word attack, timed sight word reading and decoding, and reading comprehension tasks that were measured at the end of second grade. Relative to logistic regression, improved prediction accuracy was found with classification tree analysis, with sensitivity equal to 1.0 and specificity .94. When only sound matching, rapid digit naming, oral vocabulary and initial WIF were used in the screening battery, sensitivity and specificity fell to .85 and .81, respectively. That is, without five weeks of progress monitoring, the classification accuracy of the screening battery no longer met the standard of .90 for sensitivity and specificity, but met the .75 standard.

The second approach used a weighted regression formula to determine the probability of classified risk status (i.e. "at risk" or not "at risk"). Catts, Fey, Zhang, and

Tomblin (2001) used a battery of language, early literacy, and nonverbal cognitive measures to determine the probability that kindergarten students would develop a reading disability. The criterion was a composite comprised of subtests from the WRMT-R, Gray Oral Reading Test-3, and the Diagnostic Achievement Battery-2 administered in second grade. From their findings they concluded that a five test screening battery including letter identification, sentence imitation, phonological awareness, rapid naming, and mother's education level had sufficient specificity (.91) and sensitivity (.74) for use in the initial step of the screening process. To reduce over-identification, a second wave of diagnostic testing was recommended for all students identified as "at-risk." Similar to additional weeks of progress monitoring, any additional diagnostic testing that is needed to achieve acceptable levels of sensitivity and specificity requires additional time and money.

Best practices in universal screening prescribe that screening tools be both technically adequate and efficient (Ikeda, et al., 2008). As reiterated throughout the literature, successful implementation of RTI also requires screening procedures for students at risk of reading difficulties that result in a limited number of false positives with rates of true positives that approach 100%. Under-identification of students with reading difficulties results in students who miss out on early intervention services, where over identifying students who are at risk of reading difficulties places a burden on resources. Intervention groups are either too large and therefore dilute the services received, or additional resources are required to support higher numbers of students needing intervention resulting in unsustainable tiered services. Where sufficient levels of sensitivity and specificity are achieved, lengthy screening batteries are often used that

require individual testing sessions ranging from 35 minutes to as much as four hours (Catts, et al., 2001; O'Connor & Jenkins, 1999) or the addition of at least five weeks of progress monitoring (Compton, et al., 2006). Screening procedures must be feasible for large scale implementation. Similar to over-identification, time consuming screening batteries require resources that many schools do not have, while follow-up progress monitoring is also time consuming and delays necessary intervention for students. Research is needed to improve current practices in early literacy screening in kindergarten and first grade.

Purpose

The purpose of the current study was to evaluate the extent to which early indictors of reading result in sufficient levels of accuracy to determine reading risk status when earlyReading was used for screening. The paper expands on the research literature by exploring measures that are quick and efficient to administer, and outside the narrower scope of skills assessed by DIBELS. Specifically, twelve measures of early reading were examined. Subsets of five to seven measures were given at three time points (fall, winter and spring) to students in kindergarten and first grade. Special attention was focused on exploring combinations of measures that produced sufficient levels of AUC, sensitivity and specificity estimates at each time point. The primary research questions are as follows:

1. To what extent does <u>each earlyReading</u> measure accurately predict risk status when administered in the fall, winter and spring to predict a test of broad reading achievement that was administered in the spring of kindergarten and first grade?

2. To what extent can a composite of earlyReading accurately predict risk status when administered in the fall, winter and spring to predict a test of broad reading achievement that was administered in the spring of kindergarten and first grade?

Methods

Participants

Kindergarten (N=233) and first grade students (N=180) from two school districts and six schools participated in the study. Two to three kindergarten classrooms (either half-day or full-day), and two to three first grade classrooms participated at each school. In District 1, the majority of students within the school district were White (53%), with the remaining students identified as African American (26%), Hispanic (11%), Asian (8%), or other (2%). Forty to fifty percent of students at each school received free and reduced lunch. On average, the majority of students across schools in District 2 were White (78%), with the remaining students identified as either African American (19%), or other (3%). Forty to fifty percent of students at each school received free and reduced lunch.

Measures

FAST earlyReading measures were developed as an extension of the Formative Assessment Instrumentation and Procedures for Reading (FAIP-R) project at the University of Minnesota (Christ & Ardoin, 2009). earlyReading includes 12 subtests for pre-reading and early reading skills. All subtests use standardized procedures including prescribed directions that often include a practice section with standardized response sets and timed administration.

Concepts of Print. Concepts of Print was an untimed task where the student was presented with 12 items that measure knowledge of principles related to print conventions. Students were asked to identify printed numbers, letters, shapes and sentences. Students were then asked to distinguish between words of different length that contain the same root word. For example, when presented with the words "Roll" and "Rollercoaster," the student would be asked to point to the word "Roll." The measure of performance was number correct out of 12. Concurrent validity was equal to.60 with the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) standardized total score for kindergarten in the fall. Test-retest reliability was .42 for kindergarten in the fall (N=39).

Onset Sounds. Onset Sounds was an untimed task where the student was presented with a set of four pictures and asked to point to the picture with the same onset, or beginning sound, as a word provided by the examiner. A total of 16 items were included with distinct beginning sounds across the items. A fifth set of pictures was used for practice items. After the practice and training questions, students were asked to say or point to the picture with the same beginning sound as the prompted word. For example, the examiner would say while pointing to the four pictures, "This is a fish, train, zebra, and balloon. Which one begins with /z/?" The student would then receive credit for correctly pointing to or saying "zebra." The measure of performance was the number correct out of 16. Concurrent validity was reported to equal .62 with the GRADE standardized total score for kindergarten in the fall. Test-retest reliability was .79 for kindergarten in the fall.

Letter Naming. Letter Naming was a partially timed task where the student was presented with a list of all 26 letters of the alphabet. Each letter was presented twice in both upper-and lower case. Students were asked to name each letter as quickly as possible without making mistakes. Although performance was timed for one minute students were encouraged to continue to name all the letters to provide a complete inventory of the student's letter naming accuracy. Although the measure of performance can be calculated by the total letters sounds read under timed and untimed conditions, in addition to the percent correct in one minute, the total number correct in one minute (rate score) was used as the measure of performance in the current study. Concurrent validity with the GRADE standardized total score for kindergarten in the fall was .41. Test-retest reliability was .91 for kindergarten in the fall.

Letter Sounds. Letter Sounds was very similar to Letter Names with the exception that students were asked to provide the sound of each letter instead of the name. Letters with dual sounds such as the vowels and "c" and "g" were presented at the bottom allowing for solicitation of both sounds. Dual sounds were administered during untimed conditions. Again, although the measure of performance could calculated by the total letters sounds read under timed and untimed conditions, in addition to the percent correct in one minute, the total number correct in one minute (rate score) was used as the measure of performance in the current study. Concurrent validity with the GRADE in the fall was .53 for kindergarten students. Test-retest reliability was .75 for kindergarten in the fall.

Rhyming. The format was very similar to Onset Sounds, where the student was presented with five sets of four pictures. The first set was for training and practice, with

the last four sets used for testing. For each set of pictures, three questions ask the student to say or point to the picture that rhymes with a given word. The fourth question asks the student to produce a rhyming word with a given picture. The measure of performance was the number correct out of 16. Concurrent validity with the GRADE in the fall was .58 for kindergarten students. Test-retest reliability was .74 for kindergarten in the fall.

Word Blending. Word Blending was an untimed task where the student was presented with a string of two to three sounds and asked to blend sounds to produce a word. For example, if the sounds /t/ /i/ /n/ were provided with a one second pause in between each sound, the student would have to successfully blend the sounds and produce the word "tin" to receive points. The measure of performance was the number correct out of 10. Concurrent validity with the GRADE in the fall was .22 for first grade. Test-retest reliability was .71 for kindergarten and .91 in first grade in the fall.

Word Segmenting. Word Segmenting was similar to Word Blending, except the student was provided a whole word and asked to produce the individual sounds. For example, if the word "tin" was provided, the student would respond with /t/ /i/ /n/. The measure of performance was the number of sounds correct out of 32. Concurrent validity with the GRADE in the fall was .49 for first grade. Test-retest reliability was .84 in first grade in the fall.

Nonsense Words. Nonsense Words was a timed task where the student was presented with a list of 50 nonsense words and asked to read from the list for one minute while errors were recorded. All words had the structure of consonant-vowel-consonant or vowel-consonant. The measure of performance was the number of words read correctly

in one minute. Concurrent validity estimates were not available for Nonsense Words. Test-retest reliability was .84 in first grade in the fall.

Decodable Words. Decodable Words was a timed task where the student was presented with a list of 50 decodable words and asked to read from the list for one minute while errors were recorded. All words had the structure of consonant-vowel-consonant. The measure of performance was the number of words read correctly in one minute. Concurrent validity with the GRADE in the fall was .22 for first grade. Test-retest reliability was .95 in first grade in the fall.

Sight Word 50 and 150. Sight Word was a timed task where the student was presented with a list of 50 or 150 sight words and asked to read from the list for one minute while errors were recorded. Sight Words 50 contained all 50 words on one page for kindergarten students, where Sight Words 150 contained 50 words on each of 3 pages for first grade students. The measure of performance was the number of words read correctly in one minute. Concurrent validity with the GRADE in the fall was .59 for first grade. Test-retest reliability was .97 in first grade in the fall.

Sentence Reading. Sentence Reading was a timed task where the student was presented with a series of sentences and paragraphs, and asked to read quickly without making mistakes. The number of words read correctly in one minute was recorded. One sentence was presented on a single page for the first three sentences. Four sentences were presented on the fourth page, with whole paragraphs presented on the fifth and sixth pages. In this way, only students who are successful with the limited text presented on the first three pages encounter the additional sentences on pages four through six. All sentences originate from a primer level FAIP passage (Christ & Ardoin, 2009), and

contained a related picture on each page. The measure of performance was the number of words read correctly in one minute. Test-retest reliability was .98 for first grade in the fall.

CBMReading. CBMReading is a version of CBM-R created by *FAST*. Students read aloud from a page of text while words read correct and errors are marked and recorded. The passages developed for the Grade 1 passages used in this study included 150-200 words overall in 2-5 paragraphs. Sentence length ranged from 3 to 7 words, with each paragraph containing 7 to 15 sentences. The number of words per sentence and sentences per paragraph were varied across the story to result in the appropriate total number of words. For screening purposes, students read from three passages for one minute each. The measure of performance was the median number of words read in one minute across the three passages. Concurrent validity was .89 with the Test of oral silent reading and comprehension (TOSREC), .97 with AIMSweb, and .78 with DIBELSNext for first grade students. Predictive validity was .91 with AIMSweb for first grade students after 12 weeks. Test-retest reliability was .90 and .82 for first grade from fall to winter, and fall to spring, respectively. Inter-rater reliability ranged from .83 to 1.00, with a median of .97.

Group Reading Assessment and Diagnostic Evaluation (GRADE). The GRADE was a diagnostic screening tool used to determine the reading skills children have mastered (Williams, 2001). The GRADE has eleven levels for use with students ranging from pre-kindergarten to young adulthood. Level K was administered to kindergarten students and Level 1 was administered to 1st grade students. Both assessments were group administered at the class level. The GRADE was administered in

April and May of the school year. Split-half reliability coefficients corrected by Spearman-Brown formula were reported to range from .91 to .99. Criterion related validity ranged from .76 to .90.

Level K. Level K was designed as an early reading assessment for kindergarten, early first grade, and transitional first-grade classrooms. It consists of eight required subtests and one optional subtest. The subtests include Sound Matching, Rhyming, Same and Different Words, Print Awareness, Letter Recognition, Phoneme Grapheme Correspondence, Listening Comprehension and Word Reading. Together the subtests measure phonological awareness, visual skills, early literacy skills, knowledge of print material, basic early reading skills that require both visual and auditory skills, understanding of spoken language, recognition of basic pre-primer and primer sight words and decoding simple, regular words. All nine subtests were administered to kindergarten students. The measure of performance was the overall standard score based on the raw score of all nine subtests combined.

Level 1. Level 1 was designed as an early reading assessment for kindergarten, early first grade, and transitional first-grade classrooms. It consists of five required subtests. The subtests include Word Reading, Word Meaning, Sentence Comprehension, Passage Comprehension, and Listening Comprehension. Together the subtests measure vocabulary, comprehension and oral language. All subtests were administered to 1st grade students. The measure of performance was the overall standard score based on the raw score of all five subtests combined.

Implementation Procedures

The data used in the study were obtained as a part of a pilot project of the FAST suite of assessments. School districts volunteered to administer a predetermined schedule of FAST earlyReading and CBMReading measures (see Table 1) three times per year to all students K-5 in exchange for full use of all FAST assessments at no cost. Data collection spanned one year and involved a cohort of kindergarten and first grade students.

earlyReading administration. Five to seven measures were administered to students in kindergarten and first grade at each screening period (fall, winter and spring) in the 2012-2013 academic year. On average, the screening battery took 10 to 20 minutes in kindergarten, and 8 to 12 minutes in first grade. These measures were all administered by classroom teachers and trained assistants. All teachers in District 1 were trained by district staff who attended a "train the trainer" online session. Teachers in District 2 attended a two-hour in-person training on the earlyReading measures and were observed a percentage of the time for administration integrity by the lead teacher at each school site. Estimates of administration integrity were not available.

GRADE administration. The GRADE was used as the criterion measure and group administered at the end of the school year by at least one graduate student and one additional adult. Sessions were broken into three 30 minute sessions for kindergarten, and two 45 minute sessions for first grade.

Analytic Procedure

Classification accuracy of all single measures and composites were compared within each grade level by season. Classification accuracy analysis includes the

comparison of classifications from alternative or new measures, to the classifications of an established measure or gold standard. Such analysis produces a 2 x 2 table containing true and false positives, and true and false negatives. Sensitivity and specificity are often used to summarize outcomes of classification accuracy where sensitivity is the ratio of true positives to true positives and false negatives (TP/ (TP+FN)). That is, the accuracy at which a measure (or combined measures) identifies students at-risk who later develop poor reading outcomes. Specificity is the ratio of true negatives to true negatives and false positives (TN/ (TN+FP)), or the accuracy at which a measure (or combined measures) identifies students not at-risk who later do not develop poor reading outcomes.

Area under the receiver operator curve (ROC) is also used to summarize classification accuracy. ROC's plot true-positive against false positive rates of classification for each cut-off score of the predictor. As depicted in Figure 4, the AUC serves as an indicator of predictability where .50 (the dotted line in the figure) is equal to chance, and 1.0 is equal to prefect prediction. Also depicted in Figure 4 are examples of poor, good, and very good ROC's. In the present study, AUC values of .85 and .90 were considered good and very good based on NCRtI standards and Metz (1978). For sensitivity and specificity, although .90 or above is desirable for both (especially sensitivity), the balanced approach to sensitivity and specificity used by Silberglitt & Hintze, (2005) was used in this study with values of .80 judged to be the minimum level of acceptability (Carran & Scott, 1992) and .90 judged to be desirable (Jenkins, 2003).

Silberglitt and Hintze (2005) argue that taking advantage of the highest possible diagnostic accuracy specific to a single data set rarely generalizes to other samples. Instead, a set of a priori rules were established for cut points using ROC curve analysis

where sensitivity and specificity are first brought to .70. The decision rules dictate that if a point exists on the ROC curve where sensitivity and specificity both meet .70, sensitivity is increased from that point. While still maintaining specificity of .70, sensitivity is then increased to .80 if possible. Specificity is then increased if possible, while maintaining sensitivity above .80. If both sensitivity and specificity exceed .80, this process of maximization is repeated using .90 as the next cut off.

Lastly, performance above the 30th percentile on the GRADE was used as the criterion (or standard) of risk status. The 30th percentile was used based on recommendation by Torgesen (2000). Moreover, the 30th percentile fell in between the range of criterions used in other studies, which ranged from the 15th to the 40th percentile.

Results

The means, standard deviation and range for scores on the GRADE and each of the earlyReading measures by season for kindergarten and first grade are presented in Table 2 and Table 4, respectively. The predictors included in the analyses varied by season (see Table 1). It is noted that due to missing data, Nonsense Words was excluded from spring analysis for first grade. An examination of the distribution plots for each earlyReading measure suggested floor and ceiling effects existed for some measures throughout kindergarten and first grade; however, given the extensive schedule of administration, many of these effects were expected. All measures met the assumption of linearity based on examination of the studentized residual plots, with the exception of Nonsense Words in the spring of kindergarten and Sentence Reading and CBMReading in the fall and winter of first grade, respectively. For these measures the residual values were systematically under-predicted at the high end of the distribution. After a log

transformation was applied, Nonsense Words, Sentence Reading, and CBMReading met the assumption of linearity. The log of Nonsense Words in the fall of kindergarten, and Sentence Reading and CBMReading in the fall and winter in first grade were used for all remaining analyses.

Similarly, all earlyReading measures met the assumption of independence, as the GRADE standard scores were independent of each of the earlyReading measures. The assumption of homoscedasticity was met for all earlyReading measures as the error variance did not systematically increase or decrease across any of the distributions. Lastly, for all earlyReading measures, less than 5% of the observations had residual variances that fell more than two standard deviations from the mean. For kindergarten, the correlations between the outcome measure and all required predictor variables were in the range of .41 to .59 in the fall, .46 to .62 in the winter and .43 to .54 in the spring (see Table 3). In first grade, those correlations ranged from .32 to .66 in the fall, .41 to .76 in the winter and .27 to .82 in the spring (see Table 5). Multicollinearity was evaluated using the variance inflation factor (VIF). A VIF greater than 10 was used as an indication of multicollinearity (Cohen, et al., 2003). VIF values ranged from 1.49 to 6.45 across the school year for kindergarten, and ranged from 1.39 to 6.48 across the school year for first grade; therefore multicollinearity did not appear to excessively influence the results.

Accuracy of Single Predictors

For kindergarten, the diagnostic accuracy statistics of single indicators at each time point are displayed in Table 9. Alone, no one kindergarten measure was particularly useful to identify risk status based on the end of year GRADE performance at the 30th percentile. In the fall, all ROC analysis yielded AUC values that fell below the .85

standard. Onset Sounds had the highest AUC value equal to .84. All other measures in the fall had AUC's that ranged from .68 to .79, with an average of .76. In the winter, low AUC values were again observed with the exception of Rhyming where the AUC was equal to .89. Sensitivity and specificity were both equal to .80 for Rhyming in the winter. All other measures in the winter had AUC's that ranged from .74 to .81, with an average of .79. In the spring, AUC values ranged from .73 to .80, with an average of .76. All sensitivity and specificity values were below .80.

For first grade, the diagnostic accuracy statistics of single indicators at each time point are displayed in Table 10. In the fall, AUC values for all measures with the exception of Word Segmenting exceeded .85. For Sight Words, Decodable Words, and Sentence Reading AUC values exceeded the .90 standard with values equal to .92, .92, and .93, respectively. For these measures, sensitivity and specificity ranged from .81 to .86. In the winter, AUC values for Sight Words, Decodable Words, Nonsense Words and CBMReading all exceeded .90 with values equal to .97, .96, .92 and .99, respectively. Sensitivity and specificity ranged from .84 to .96. Sensitivity and specificity were highest for CBMReading where sensitivity equaled .93 and specificity equaled .96. In the spring, similar results were observed, where AUC values for Sight Words, Decodable Words and CBMReading equaled .96, .96, and .98, respectively. Sensitivity and specificity values were equal to .91 and .89 for Sight Words, and .87 and .85 for Decodable Word, respectively, with higher levels of sensitivity and specificity observed for CBMReading (.91 and .96, respectively).

Accuracy of Composites

Using composite formulas, diagnostic accuracy statistics were improved over single measures in kindergarten (see Table 11). AUC values for the full composite (i.e. all required indicators as part of the study) were equal to .86, .91, and .82 for the fall, winter and spring, respectively. When the total number of indicators included in the composite was reduced to four measures, AUC values were comparable to the full model at all time points in kindergarten. When using two and three measures, slight decreases in AUC values were observed compared to the full model. In the fall, AUC values for the top three subsets of measures for composites containing two and three measures ranged from .82 to .86, and .85 to .86, respectively. In the winter, AUC values ranged from .85 to .90 and .90 to .91 for two and three measure subsets, where in the spring AUC values ranged from .81 to .82 and .81 to .83, respectively. Sensitivity and specificity for all subsets in the winter were equal to or greater than .80. In the fall and spring, lower values in the .70 to .80 range were observed for sensitivity and specificity.

Using composite formulas in first grade, diagnostic accuracy statistics were improved over single indicators only in the fall (see Table 12). In the fall, where the AUC value for the full composite was equal to .95, AUC values ranged from .93 to .95 for the top three subsets containing two measures. For subsets containing two to three measures, sensitivity and specificity values ranged from .80 to .86. In the winter and spring, AUC values for CBMReading alone were comparable to using all measures combined. AUC values were equal to .99 and .98 for winter and spring, respectively. All sensitivity and specificity values approached or exceeded .90 for all subsets in the winter and spring.

Discussion

This study examined the classification accuracy of early reading indicators among kindergarten and first grade students. In kindergarten the minimum accepted level of diagnostic accuracy estimates were achieved in the fall and winter when using composites with two efficient early reading measures within the universal screening battery. In the spring of kindergarten, diagnostic accuracy estimates of composites using two measures approximated, but did not exceed these standards. In first grade, diagnostic accuracy estimates using composites with two measures in the fall met the minimum standard. By winter, desirable diagnostic accuracy estimates were observed when using only CBMReading. Compared with the concurrent and predictive validity estimates observed by Monaghen (2014) using the same data set, diagnostic accuracy estimates obtained in this study indicate the use of fewer measures within screening battery composites.

In kindergarten, results were generally consistent with previous research in that AUC, sensitivity and specificity estimates of single indicators continued to fall short of acceptable standards for even low-stakes decisions (Catts, et al., 2001; Compton, et al., 2006; Hintze, et al., 2003; Johnson, et al., 2010; Nelson, 2008; Riedel, 2007). In the current study, AUC standards of .85 and .90 were interpreted as good and very good, respectively. For sensitivity and specificity, minimum standards were equal to .80 or above, with .90 or above designated as desirable. For kindergarten, Rhyming in the winter was the only measure that met standards for good AUC values with minimal levels of sensitivity and specificity. Other earlyReading indicators met the lower standard of .75 for both sensitivity and specificity used by Hintze et al. (2003). These included Onset

Sounds in the fall, Letter Sounds and Word Segmenting in the winter, and Decodable Words and Nonsense Words in the spring. AUC values were also above .75 for those measures. This suggests that universal screening measures for kindergarten are improving, but continue to fall short of higher standards. This is not surprising as the individual indicators used in kindergarten are direct measures of relatively constrained skills, which are unlikely to predict and classify students in the same way as broad measures of reading achievement.

The less than desirable results of single measures in kindergarten universal screening provide further support that multiple early reading measures are required to achieve adequate levels of diagnostic accuracy estimates (Johnson, et al., 2010; Johnson, et al., 2009). Using weighted composite formulas based on multiple regressions, the top three subsets of indicators were identified using each of two to five earlyReading measures. In kindergarten, diagnostic accuracy estimates using only two measures were comparable to using the full battery of measures (six to seven measures) across fall, winter, and spring; however, desirable levels of diagnostic accuracy estimates were only observed in the fall and winter. That is, the use of multiple measures for universal screening met the standard of .85 for AUC and .80 for sensitivity and specificity in the fall and winter, but not the spring of kindergarten.

In comparison to composites identified by Monaghen (2014), based on criterion validity, composites chosen in this study place a heavier emphasis on efficiency and technical adequacy of universal screening. Given the use of only two measures across kindergarten, fewer early reading domains (i.e., concepts of print, phonemic awareness, alphabetic principle or decoding) are measured at each time point. The use of additional

measures within the composite would increase measurement across the domains, but increase administration time with little value added to the diagnostic accuracy of universal screening decisions in kindergarten. A notable subset of measures in the fall included of Rhyming and Onset Sounds (AUC equal to .86, sensitivity and specificity equal to .79 and .83, respectively). Other notable combinations included Rhyming and Word Segmenting (AUC equal to .90, sensitivity and specificity both equal to .82, respectively) in the winter, and Rhyming and Decodable Words (AUC equal to .82, sensitivity and specificity equal to .80 and .77, respectively) in the spring (see Table 13).

Another notable finding was the difference in observed estimates for the same measure across the fall, winter and spring of kindergarten. For example, in the current study where Onset Sounds was useful in discriminating between students in the fall who later met the 30th percentile criterion on the GRADE at the end of the school year, the diagnostic accuracy of Onset Sounds diminished from fall to winter. This is in contrast to other measures, such as Letter Sounds and Rhyming where estimates improved from fall to winter, before declining again in the spring. The differences across time in observed estimates of AUC, sensitivity and specificity highlight the constrained and changing nature of reading skills in the early grades (Paris, 2005). Unlike the later elementary grades, where a single reading measure is useful in the fall, winter and spring of a given school year (Johnson, Jenkins, & Petscher, 2010; Wayman, Wallace, Wiley, Ticha, & Espin, 2007), in kindergarten the rate of growth in reading skills may be too rapid for a consistent universal screening measure across the school year. The results provide support for use of different universal screening batteries across the kindergarten school year. Use of different measures is also aligned with the development and the scope and

sequence of reading instruction of early reading skills (Chall, 1983; Simmons & Kame'enui, 1999).

By first grade, use of Sentence Reading or CBMReading alone were judged to approximate performance of multiple measure composites used in previous research (Catts, et al., 2001; Compton, et al., 2006). Diagnostic accuracy estimates of Sentence Reading met the minimum standard in the fall (AUC equal to .93, sensitivity and specificity equal to .89 and .81, respectively). In the winter and spring, desirable standards were met with CBMReading in the winter (AUC equal to .99, sensitivity and specificity equal to .93 and .96, respectively), and spring (AUC equal to .98, sensitivity and specificity equal to .91 and .96, respectively). Desirable AUC values were observed with use of Sight Words or Decodable Words across the school year, but sensitivity and specificity estimates continued to fall short of desirable standards. When using multiple measures within a composite (Word Segmenting and Sentence Reading), a two percent increase in AUC was observed over use of a single indicator only in the fall, but again similar increases were not observed for sensitivity and specificity. Similarly in the winter and spring, diagnostic accuracy estimates were not increased with use of composites for first grade students. That is, performance of CBMReading alone was found comparable to multiple measure composites in the winter and spring of first grade. Composites identified in first grade with diagnostic accuracy (see Table 14) were directly comparable to composites identified using criterion related validity by Monaghen (2014), with use of Word Segmenting and Sentence Reading in fall and CBMReading alone in the winter and spring.

Implications

In kindergarten, use of single indicators for universal screening is not supported based on results from this study. As suggested by Johnson et al. (2010), such efforts are probably futile due to the constrained skills of early reading (Paris, 2005); however, use of multiple measures based on weighted composite formulas show promise for reaching acceptable standards of diagnostic accuracy. The FAST composites identified in this study allow educators to make low-stakes decisions about which students are in need of tier II services. Likewise, unlike the use of group administered tests of broad reading achievement like the GRADE, the use of FAST earlyReading measures provide teachers with classwide data on specific reading skills related to instruction and are directly linked to progress monitoring of these early reading skills.

Universal screening tools that are reliable, valid, and efficient are essential to the success of RTI. Continued research should focus on alternate measures and varying combinations of measures that are aligned with kindergarten reading instruction. Where in the current study the inclusion of the FAST Concepts of Print measure in the fall screening battery helped to eliminate floor effects in early kindergarten that were previously observed in other studies (Catts, et al., 2001; McAlenney & Coyne, 2011), the identification of alternate measures in the future may further support the direct placement of students within tier II interventions. In contrast to lengthy screening batteries followed up by weeks of progress monitoring, more efficient universal screening batteries are emerging as a viable alternative for kindergarten students.

In first grade, use of single measures produced sufficient diagnostic accuracy estimates for use within low-stakes decisions such as universal screening. With false

positive and negative rates that improve upon those observed in past research (Johnson, et al., 2009; Riedel, 2007), support for use of only Sentence Reading or CBMReading as a universal screener in first grade is further established. For example, in the current study when sensitivity for Sentence Reading in the fall of first grade is brought to the .90 level used by Johnson et al., sensitivity and specificity equal .95 and .76, respectively. This is in comparison to observed sensitivity and specificity estimates of .90 and .65, respectively, for oral reading fluency in the fall of first grade found in previous research (Johnson, et al., 2009). The improvements are again attributed to the improvements in the tools used in this study. The FAST suite of assessment used strict writing specifications for first grade level passages (i.e. Sentence Reading and CBMReading), and tested passages on hundreds of students across three phases of data collection to ensure passage readability. Similarly, the strategies used in the format of Sentence Reading likely encourage young students to read more words based on the presentation of one sentence and a picture per page.

Limitations and Future Directions

Although these results are encouraging, the limitations must be taken into consideration. First, it is important to remember that research focused on the diagnostic accuracy of early reading measures for universal screening can still be considered infantile. Future research is needed to continue to explore alternate measures and potential improvements to existing measures in order to further increase levels of AUC, sensitivity and specificity. Developments in research thus far are encouraging for use of multiple measures in kindergarten screening, and the use of oral reading fluency from the beginning of first grade. Continued research is encouraged.

Secondly, the sample from which the diagnostic accuracy statistics were estimated may limit generalization of results. All schools were from the Midwest, with classroom participation based on schools and teachers who volunteered. The differences in training models used by the two districts involved this study may also contribute to differences in administration. The composites evaluated were also identified through data analysis methods that were highly data-driven and therefore subject to sampling error. Future research should focus on the cross-validation of results. Similarly, despite efforts to ensure adherence of the test administration schedule, some teachers failed to administer required measures at each time point, namely Nonsense Words. For example, in the winter of kindergarten only 104 students out of 233 and 131 out of 188 first grade students in the spring were administered the Nonsense Word measure. While missing data is a common problem in research with several methods of imputation available, missing data rates ranging from 30 to 50 percent as observed in this study, have the potential to significantly skew results (Bennett, 2001; Peng, Harwell, Liou, & Ehman, 2007 ; Schafer, 1999). For this reason, Nonsense Words was not included in analysis for those specific time points.

Outside of missing data, a second limitation of the test administration schedule involves the absence of Sentence Reading as a required measure in the spring of kindergarten. Patterns of AUC values across the kindergarten year highlight that by spring, measures of word decoding (i.e., Nonsense Words and Decodable Words) are relatively better at discriminating between students than measures of phonological awareness. Given that, and that more desirable diagnostic accuracy estimates were

observed for Sentence Reading in the fall of first grade, it is hypothesized that the use of Sentence Reading in the spring of kindergarten might also produce desirable estimates.

The criterion used in this study also influences results. While it is the assumption that the standard is 100% accurate, use of a different measure and criterion performance level might produce different results. For example, as Jenkins, Hudson, and Johnson (2007) highlight, the outcome criterion used in similar studies varies drastically with use of the Woodcock Johnson Test of Achievement – Revised, Woodcock Reading Master Test, Stanford Achievement Test, 10th ed., CBM-R, or the double discrepancy commonly used in kindergarten and first grade research. Likewise, the criterion performance level, or the method for choosing the cut points, can influence the levels of sensitivity and specificity. Where some select cut-points based on a pre-specified level of sensitivity (i.e. 75, 90, or 95 percent; Foorman, et al., 1998; O'Connor & Jenkins, 1999), others select the highest sensitivity associated with a pre-specified cut point (D. L. Speece & Case, 2001), and still others choose a balanced approach with preference given to sensitivity (Silberglitt & Hintze, 2005). As described by Speece (2005), the decision between choosing under-identification and over-identification is a matter of picking your poison.

Conclusion

The results from both grade levels support the importance of diagnostic accuracy studies in addition to studies focused on the criterion-related validity. In support of Jenkins (2003) and Nelson's (2008) position that diagnostic accuracy is an essential component to establishing the utility of early reading screening measures, the current results highlight the added value of classification validity. Where results from Monaghen (2014) indicated that the use of three to four measures in kindergarten, two to three

measures in the fall of first grade, and one to two measures in the winter and spring of first grade, were necessary to meet standards of concurrent and predictive validity for universal screening, results of the current study indicated fewer measures were needed to meet minimum and desired levels of diagnostic accuracy estimates. With more time efficient and cost effective screening procedures that maximize the number of correct classifications, implications include improved reading outcomes for students.

Chapter 5

The purpose of the two studies included in this paper was to examine the concurrent and predictive validity and diagnostic accuracy of single and multiple reading measures for universal screening in kindergarten and first grade. Efforts and emphasis on prevention, early intervention, and RTI rely on high quality measures. Moreover, prevention and early intervention require a high quality screening system for emergent readers in kindergarten and first grade. The present studies establish the benefits of FAST earlyReading within an RTI framework to contribute to early identification and prevention of reading difficulties in students through effective universal screening in the early grades. The studies further demonstrate the improvements of FAST earlyReading over previous measures of early reading. Use of FAST earlyReading composites based on regression formulas improve the technical adequacy of the measures and create ease of interpretation for educators to identify students for tier II intervention.

In Study 1, the concurrent and predictive validity of each predictor, and varying composites were examined. Based on NCRtI criteria of $r \geq .70$ ($R^2 = .49$), use of single measures for universal screening was not supported at any time point in kindergarten. The use of composites consisting of three to four measures significantly improved estimates, but continued to fall short of criteria in the fall of kindergarten. By winter of kindergarten, use of composites with four measures were observed to meet criteria. Such composites were comparable across the winter and spring to the full screening battery consisting of six to seven measures. In first grade, Sentence Reading and CBMReading alone were observed to meet minimum criteria for use as a universal screener. When composites were used with two to three measures in the fall, validity estimates were

observed to increase to more desirable levels. By winter, smaller increases in validity estimates were observed for composites compared with Sentence Reading and CBMReading alone. Overall recommendations based on Study 1 include use of composites consisting of three to four FAST earlyReading measures during kindergarten, one to two measures in the fall of first grade, and CBMReading in the winter and spring of first grade.

In Study 2, the diagnostic accuracy of each predictor, and varying composites were examined. AUC values of .85 and .90 were considered good and very good based on NCRtI standards and Metz (1978). The minimum level of acceptability for values of sensitivity and specificity were judged to be equal to or above .80 with values equal to or above .90 judged to be desirable (Jenkins, 2003). In kindergarten the minimum level of diagnostic accuracy estimates were achieved in the fall and winter when using two early reading measures within the universal screening battery. In the spring of kindergarten, diagnostic accuracy estimates when using two measures approached these standards, but ultimately fell short. In first grade, diagnostic accuracy estimates using two measures in the fall met the minimum standard. By winter, desirable diagnostic accuracy estimates were observed when using only CBMReading. Overall recommendations based on Study 2 include use of composites consisting of two FAST earlyReading measures during kindergarten, one to two measures in the fall of first grade, and CBMReading in the winter and spring of first grade.

Together, the concurrent and predictive validity and diagnostic accuracy of FAST earlyReading measures and composites help establish a case for validity of these measures for use as universal screeners of reading in kindergarten and first grade.

Estimates observed of the concurrent and predictive validity and diagnostic accuracy of composites improve upon those observed in previous studies and support use of composites as universal screeners in kindergarten and use of only CBMReading by the end of first grade. The comparison of recommendations from both studies indicates fewer measures are required to meet standards of diagnostic accuracy than standards of criterion-related validity, especially in kindergarten.

Difference in recommendations from Study 1 and Study 2 may stem from standards of criterion-related validity that are higher than necessary for decisions related to universal screening. Good et al. (2011), and Hopkins (2002) indicate that correlations equal to or above .70 are "strong" (p.94) or "very large, very high, huge," where correlations between .50 and .70 are "moderate-strong" (p.94) or "large, high, major." Using a lower criterion-related validity standard of .50 (R^2 = .25) would alter the interpretation of estimates observed in Study 1, where some single measures would be judged as sufficient for low-stakes decisions. Using a lower standard would also align results from the current studies with those from previous research which found moderate to strong criterion-related validity estimates and only moderate estimates of diagnostic accuracy (Nelson, 2008). Future research must consider the standards at which estimates are compared.

Implications

Critics of RTI and universal screening suggest that accurate identification of "at-risk" beginning readers requires 5 to 8 weeks of progress monitoring or a second wave of screening in order to sufficiently reduce rates of false-positives (Catts, et al., 2001; D. Fuchs, et al., 2012; Nelson, 2008; O'Connor & Jenkins, 1999). The present studies

provide evidence that FAST earlyReading composites created using regression based formulas efficiently identify "at-risk" students with only 10 minutes per student. These composites further contribute to the ease of interpretation for educators with use of single scores that summarize the combination of assessments most useful in the fall, winter, and spring of a given school year. FAST earlyReading measures and use of composites are increasing the capacity of universal screening to be used for accurate classifications of students "at-risk" of reading difficulties in the younger grades. While use of single indicators in kindergarten is not supported, universal screeners that use FAST earlyReading composite consisting of two to four measures are supported in these studies. Contrary to other recommendations, results from the current study also support that universal screening in kindergarten can occur within the early weeks of reading instruction. By first grade, use of single indicators was recommended.

As universal screening is the first step within the RTI process, successful RTI implementation hinges on accurate identification of "at-risk" students especially for kindergarten and first grade students. Intervention is most effective when provided early. If accurate identification of students with reading difficulties does not occur until the later elementary grades, it may be too late to successfully remediate such difficulties for many students (Adams, 1994; Foorman, Francis, Shaywitz, Shaywitz, & Fletcher, 1997). Evidence from the current studies provide evidence for use of FAST earlyReading to contribute to effective RTI implementation. Continued research should focus on the improvement of early reading universal screening measures and the replication of similar studies using FAST earlyReading measures to further establish screening composites with high levels of classification accuracy throughout kindergarten and first grade.

Limitations

Limitations from both studies include the sample from which the validity estimates were obtained as generalization of results may be limited. Future research should focus on the cross-validation of results. Test administration schedule and adherence were also met with limitations. Sentence Reading was not administered in the spring of kindergarten, and many teacher participants did not administer Nonsense Words as indicated in the schedule leading to large percentages of missing data. The inclusion of alternate measures not included in the *FAST* suite of assessments might also contribute to higher validity estimates in the future. Likewise, the criterion used in this study also influences results, where use of a different measure or criterion performance level might produce different results. Lastly, unidimensionality was not found across kindergarten and first grade. As the RTI framework is used to identify dual discrepancies in level and rate of reading achievement, the comparison of composite scores across time would be useful to measure student growth; however, the break unidimensionality prevents a direct comparison at this time.

Tables

Table 1. Screening administration schedule for kindergarten and first grade for fall, winter and spring.

Assessments	Kindergarten			1st		
	Fall	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
1) Concepts of Print	X	O	O			
2) Onset Sound	X	X	O			
3) Letter Naming Fluency	X	X	X			
4) Letter Sound Fluency	X	X	X			
5) Rhyming	X	O	O			
6) Word Blending Fluency	X	X	X	X	X	X
7) Word Segmenting Fluency	O	X	X	X	X	X
8) Nonsense Word Fluency	O	X	X	X	X	X
9) Sight Word Fluency		X	X	X	X	X
10) Decodable Word Fluency		O	X	X	X	X
11) Sentence Reading				X	CBM-R	CBM-R

Notes: O = Optional; X = Required

Table 2. Descriptive statistics for earlyReading measures and GRADE standard scores in a sample of kindergarten students across three seasons.

	N	Mean	SD	Range	Max Possible	Skew	Kurtosis
GRADE SS	218	105.32	15.02	66 – 135	135	0.09	-0.31
FALL							
Concepts of Print	229	8.41	2.41	0 – 12	12	-0.72	0.58
Onset Sounds	229	12.28	4.17	0 – 16	16	-1.15	0.33
Letter Name (rate)	229	28.67	15.56	0 – 52	52+	-0.28	-0.97
Letter Sound (rate)	229	15.58	11.71	0 – 59	52+	0.59	-0.10
Rhyming	229	9.48	4.98	0 – 16	16	-0.37	-1.06
Word Blending	229	3.18	3.68	0 – 10	10	0.63	-1.23
Word Segmenting	90	6.23	9.68	0-34	34	1.38	0.53
WINTER							
Concepts of Print	57	9.42	2.15	5-12	12	-0.25	-1.03
Onset Sounds	215	15.06	2.04	2-16	16	-3.24	12.42
Letter Name (rate)	209	39.55	13.37	1-52	52+	-0.98	0.01
Letter Sound (rate)	209	29.19	13.77	0-66	52+	0.24	0.39
Rhyming	223	12.00	4.63	0-16	16	-1.23	0.46
Word Blending	226	6.75	3.36	0-10	10	-0.85	-0.67
Word Segmenting	227	19.29	12.89	0-34	34	-0.44	-1.45
Nonsense Words	104	13.16	8.56	0-49	50+	1.94	5.13
SPRING							
Onset Sounds	154	15.70	1.61	0-16	16	-8.13	71.28
Letter Name (rate)	229	54.80	18.40	2-124.80	52+	0.07	0.70
Letter Sound (rate)	229	43.10	15.51	0-86.09	52+	-0.18	0.04
Rhyming	228	14.28	3.15	1-16	16	-2.29	4.84
Word Blending	228	9.14	1.79	0-10	10	-3.01	9.93
Word Segmenting	227	30.11	6.17	0-34	34	-2.63	7.69
Decodable Words	227	16.04	15.15	0-85.71	50+	1.49	2.28
Nonsense Words	228	13.16	10.71	0-50	50+	1.42	2.21
Sight Words 50	226	44.29	29.38	0-130.43	50+	0.38	-0.63

Table 3. Correlation among predictor and outcome variable for kindergarten.

Measure	1	2	3	4	5	6	7	8	9	10	11
FALL											
1. GRADE.SS	--	214	214	214	214	214	214	--	--	--	--
2. Concepts	0.50	--	229	229	229	229	229	--	--	--	--
3. Onset Sounds	0.56	0.43	--	229	229	229	229	--	--	--	--
4. LetterName	0.47	0.48	0.54	--	229	229	229	--	--	--	--
5. LetterSound	0.44	0.43	0.52	0.75	--	229	229	--	--	--	--
6. Rhyming	0.59	0.51	0.56	0.43	0.38	--	229	--	--	--	--
7. Word Blending	0.41	0.43	0.43	0.45	0.49	0.37	--	--	--	--	--
WINTER											
1. GRADE.SS	--	50	201	195	195	209	212	213	95	--	--
2. Concepts	0.57	--	57	57	57	56	56	56	16	--	--
3. Onset Sounds	0.47	0.60	--	195	195	210	213	214	94	--	--
4. LetterName	0.49	0.37	0.43	--	209	207	207	207	89	--	--
5. LetterSound	0.57	0.49	0.48	0.74	--	207	207	207	89	--	--
6. Rhyming	0.62	0.61	0.42	0.46	0.48	--	223	223	103	--	--
7. Word Blending	0.50	0.46	0.52	0.48	0.53	0.49	--	226	104	--	--
8. Word Segmenting	0.57	0.46	0.39	0.41	0.50	0.45	0.53	--	104	--	--
9. Nonsense Words	0.46	0.59	0.32	0.46	0.61	0.31	0.47	0.47	--	--	--
SPRING											
1. GRADE.SS	--	--	141	215	215	215	214	214	215	216	214
2. Concepts	--	--	--	--	--	--	--	--	--	--	--
3. Onset Sounds	0.17	--	--	152	152	151	150	150	149	150	148
4. LetterName	0.43	--	0.19	--	229	227	226	226	226	226	225
5. LetterSound	0.43	--	0.26	0.66	--	227	226	226	226	226	225
6. Rhyming	0.54	--	0.09	0.30	0.35	--	227	227	225	225	225
7. Word Blending	0.48	--	0.23	0.31	0.33	0.46	--	227	224	224	224
8. Word Segmenting	0.50	--	0.22	0.38	0.39	0.52	0.71	--	224	224	224
9. Nonsense Words	0.52	--	0.14	0.59	0.48	0.28	0.28	0.27	--	227	226
10. Decodable Words	0.52	--	0.15	0.59	0.51	0.29	0.31	0.30	0.90	--	226
11. Sight Words 50	0.43	--	0.17	0.71	0.47	0.32	0.34	0.38	0.74	0.68	--

Note: Correlations are below diagonal. N's are above diagonal.

Table 4. Descriptive statistics for earlyReading measures and GRADE standard scores in a sample of first grade students.

	N	Mean	SD	Range	Max Possible	Skew	Kurtosis
GRADE SS	233	112.30	15.01	65 – 145	145	-0.22	-0.28
FALL							
Word Blending	175	7.43	2.84	0 – 10	10	-1.31	0.75
Word Segmenting	175	26.84	6.96	0 – 34	34	-1.57	2.47
Decodable Words	175	14.41	14.67	0 – 49	50+	1.26	0.39
Nonsense Words	175	10.91	9.04	0 – 48	50+	1.38	1.97
Sight Words	175	34.61	23.48	0 – 97	150+	0.34	-0.65
Sentence Reading	175	43.88	36.69	1 – 192	192	1.74	3.50
WINTER							
Word Blending	164	8.80	1.90	0 – 10	10	-2.75	8.10
Word Segmenting	163	30.13	4.94	1 – 34	34	-2.70	9.70
Decodable Words	165	26.15	14.83	0 – 50	50+	0.23	-1.14
Nonsense Words	161	20.67	12.72	0 – 50	50+	0.83	-0.88
Sight Words	162	60.16	25.24	3 – 127	150+	-0.33	-0.17
CBM-R (median of 3)	179	75.48	42.55	6-218.30	150+	0.86	0.56
SPRING							
Word Blending	168	9.30	1.38	2 – 10	10	-3.00	10.64
Word Segmenting	168	30.91	4.40	1 – 34	34	-2.95	13.41
Decodable Words	182	40.80	21.12	1 – 94.29	50+	0.45	-0.39
Nonsense Words	127	24.26	14.31	0 – 66.67	50+	0.87	-0.33
Sight Words	169	77.03	21.93	13 – 141	150+	-0.22	0.13
CBM-R (median of 3)	188	104.00	43.19	20-233.20	150+	0.45	0.09

Table 5. Correlation among predictor and outcome variable for first grade.

Measure	1	2	3	4	5	6	7
FALL							
1. GRADE.SS	--	173	173	173	173	173	173
2. Word Blending	0.56	--	175	175	175	175	175
3. Word Segmenting	0.32	0.56	--	175	175	175	175
4. Decodable Words	0.58	0.35	0.20	--	175	175	175
5. Nonsense Words	0.60	0.40	0.24	0.82	--	175	175
6. Sight Words	0.66	0.40	0.20	0.80	0.72	--	175
7. Sentence Reading	0.63	0.36	0.15	0.88	0.76	0.83	--
WINTER							
1. GRADE.SS	--	162	161	161	159	160	177
2. Word Blending	0.44	--	163	162	159	161	164
3. Word Segmenting	0.41	0.61	--	162	159	161	163
4. Decodable Words	0.72	0.41	0.37	--	159	162	163
5. Nonsense Words	0.67	0.41	0.36	0.88	--	158	161
6. Sight Words	0.74	0.44	0.43	0.81	0.73	--	162
7. CBM-Reading (median)	0.76	0.31	0.25	0.83	0.79	0.80	--
SPRING							
1. GRADE.SS	--	166	166	180	125	167	182
2. Word Blending	0.35	--	168	168	126	168	168
3. Word Segmenting	0.27	0.52	--	168	126	168	168
4. Decodable Words	0.65	0.36	0.27	--	127	169	182
5. Nonsense Words	0.68	0.37	0.25	0.83	--	127	127
6. Sight Words	0.65	0.38	0.39	0.75	0.64	--	169
7. CBM-Reading (median)	0.82	0.31	0.24	0.76	0.75	0.74	--

Note: Correlations are below diagonal. N's are above diagonal

Table 6. Example fitted regression models for predicting GRADE standard scores at each time point in kindergarten.

	Parameter Estimates			
	Fall	Winter	Spring	
	Model A	Model B	Model C	Model D
Intercept	72.99*** (3.23)	77.16*** (2.20)	58.82*** (4.22)	59.47*** (4.32)
Concepts of Print	1.25** (0.39)	-	-	-
Onset Sounds	0.91*** (0.24)	-	-	-
Letter Sounds	0.13 . (0.08)	0.25*** (0.07)	-	0.06 (0.06)
Word Segmenting	-	0.35*** (0.07)	0.49** (0.16)	-
Rhyming	0.92*** (0.20)	1.24*** (0.19)	1.38*** (0.28)	1.39*** (0.29)
Nonsense Words	-	-	3.71** (1.41)	0.60*** (0.15)
Decodable Words	-	-	0.22** (0.07)	0.34*** (0.06)
R^2	0.46	0.54	0.49	0.48

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 7. Domains represented within the top three kindergarten composites using one to five predictors.

Fall					Winter					Spring				
Measures	CP	PA	AP	DC	Measures	CP	PA	AP	DC	Measures	CP	PA	AP	DC
R		*			LS			*		DW				*
O		*			WS		*			R		*		
C	*				R		*			NW			*[2]	*[2]
O + R		*			O + R		*			WS + DW				*
C+R	*	*			LS +R		*	*		R + DW		*		*
LN+R		*[1]			WS + R		*			R + NW		*	*[2]	*[2]
R+C+O	*	*			R + LN + WS		*[1]			R + WS+ DW		*		*
R+LN+O		*[1]			R + WS + O		*			R + WB + DW		*		*
R+LS+O		*	*		R + LS + WS		*	*		R + WB + NW		*	*[2]	*[2]
C+O+LN+R	*	*[1]			O+LN+R+WS		*[1]			R+WB+WS+DW		*		*
C+O+R+WB	*	*			LS+R+WB+WS		*	*		R+WS+NW+DW		*	*[2]	*[2]
C+O+LS+R	*	*	*		O+LS+R+WS		*	*		R+WB+NW+DW		*	*[2]	*[2]
C+O+LN+R+WB	*	*[1]			LN+LS+R+WB+WS		*[1]	*		R+WS+NW+SW+DW		*	*[2]	*[2]
C+O+LS+R+WB	*	*	*		O+LN+LS+R+WS		*[1]	*		R+WB+NW+SW+DW		*	*[2]	*[2]
C+O+LN+LS+R	*	*[1]	*		O+LS+R+WB+WS		*	*		R+WB+WS+NW+DW		*	*[2]	*[2]
Full Model	*	*[1]	*		Full Model		*[1]	*		Full Model		*	*[2]	*[2]

* indicates a measure of domain is included in composites; CP = Domain of Concepts of Print; PA = Domain of Phonemic Awareness; AP = Domain of Alphabetic Principle; DC = Domain of Decoding; [1] Letter Names also included. [2]Nonsense Words is included under Alphabetic Principle and Decoding

Table 8. Example fitted regression models for predicting GRADE standard scores at each time point in first grade.

	Parameter Estimates (SE)		
	Fall	Winter	Spring
	Model E	Model F	Model G
Intercept	56.13*** (4.40)	27.44*** (4.54)	79.75*** (1.80)
Word Segmenting	0.39** (0.12)		
Sentence Reading	13.24*** (1.03)		
CBMReading		20.49*** (1.08)	0.31*** (0.02)
R^2	0.55	0.67	0.68

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 9. Kindergarten Diagnostic Accuracy for fall, winter and spring using single predictors.

Measure	Season	Cut	Threshold	AUC	Sens	Spec	Accuracy	PPV	NPV	tp	tn	fp	fn
Concepts	Fall	30	8	0.77	0.62	0.85	0.67	0.35	0.95	35	108	65	6
Onset	Fall	30	11	0.84	0.79	0.78	0.79	0.46	0.94	32	136	37	9
LN	Fall	30	24	0.76	0.68	0.76	0.70	0.36	0.92	31	118	55	10
LS	Fall	30	12	0.72	0.58	0.76	0.61	0.30	0.91	31	100	73	10
RHYM	Fall	30	7	0.79	0.74	0.73	0.74	0.40	0.92	30	128	45	11
WB	Fall	30	1	0.68	0.57	0.80	0.61	0.31	0.92	33	98	75	8
Onset	Winter	30	15	0.74	0.74	0.64	0.72	0.37	0.90	25	120	42	14
LN	Winter	30	39	0.77	0.72	0.70	0.71	0.39	0.90	28	111	44	12
LS	Winter	30	25	0.81	0.76	0.75	0.76	0.45	0.92	30	118	37	10
RHYM	Winter	30	11	0.89	0.80	0.80	0.80	0.49	0.94	32	136	33	8
WB	Winter	30	6	0.76	0.75	0.72	0.75	0.40	0.92	29	129	43	11
WS	Winter	30	11	0.79	0.76	0.78	0.77	0.44	0.94	32	131	41	9
LN	Spring	30	50	0.76	0.72	0.75	0.73	0.38	0.93	30	126	49	10
LS	Spring	30	38	0.76	0.67	0.7	0.68	0.33	0.91	28	118	57	12
RHYM	Spring	30	15	0.77	0.66	0.75	0.67	0.33	0.92	30	115	60	10
WB	Spring	30	9	0.73	0.74	0.64	0.72	0.36	0.90	25	130	45	14
WS	Spring	30	31	0.76	0.64	0.72	0.65	0.31	0.91	28	112	63	11
SW	Spring	30	29	0.75	0.74	0.71	0.73	0.39	0.91	29	128	45	12
DW	Spring	30	7	0.80	0.76	0.76	0.76	0.42	0.93	31	132	42	10
NW	Spring	30	8	0.78	0.75	0.76	0.75	0.42	0.93	32	130	44	10

Note: Cut = 30th percentile cut point on the GRADE; Threshold = cut score identified on earlyReading measure; Concepts = Concepts of Print; AUC = Area under the Curve; Sens = Sensitivity; Spec = Specificity; Accuracy =Overall Classification Accuracy; PPV = Positive Predictive Value; NPV= Negative Predictive Value; TP = True Positives; TN = True Negative; FP = False Positives; FN = False Negatives; Onset = Onset Sounds; LN = Letter Names; LS = Letter Sounds; RHYM = Rhyming; WB = Word Blending; WS = Word Segmenting; SW = Sight Words 50 ; DW = Decodable Words; NW = Nonsense Words.

Table 10. First grade Diagnostic Accuracy for fall, winter and spring using single predictors.

Measure	Season	Cut	Threshold	AUC	Sens	Spec	Accuracy	PPV	NPV	tp	tn	fp	fn
WB	Fall	30	6	0.86	0.84	0.76	0.83	0.40	0.96	16	128	24	5
WS	Fall	30	25	0.80	0.77	0.76	0.77	0.31	0.96	16	117	35	5
SW	Fall	30	12	0.92	0.83	0.81	0.83	0.40	0.97	17	126	26	4
DW	Fall	30	3	0.92	0.86	0.81	0.86	0.45	0.97	17	131	21	4
NW	Fall	30	5	0.88	0.79	0.86	0.80	0.36	0.98	18	120	32	3
SR	Fall	30	16	0.93	0.89	0.81	0.88	0.50	0.97	17	135	17	4
WB	Winter	30	9	0.83	0.50	0.96	0.57	0.24	0.99	22	70	69	1
WS	Winter	30	30	0.81	0.67	0.74	0.68	0.27	0.94	17	93	45	6
SW	Winter	30	44	0.97	0.88	0.91	0.88	0.55	0.98	21	120	17	2
DW	Winter	30	15	0.96	0.84	0.87	0.84	0.48	0.97	20	116	22	3
NW	Winter	30	12	0.92	0.82	0.87	0.82	0.44	0.97	20	111	25	3
CBM	Winter	30	38	0.99	0.93	0.96	0.93	0.68	0.99	23	142	11	1
WB	Spring	30	9	0.66	0.69	0.52	0.66	0.25	0.88	14	96	43	13
WS	Spring	30	32	0.69	0.53	0.74	0.56	0.23	0.91	20	73	66	7
SW	Spring	30	61	0.96	0.91	0.89	0.90	0.65	0.98	24	127	13	3
DW	Spring	30	24	0.96	0.87	0.85	0.87	0.53	0.97	23	133	20	4
NW	Spring	30	16	0.89	0.82	0.83	0.82	0.53	0.95	20	83	18	4
CBM	Spring	30	67	0.98	0.91	0.96	0.92	0.65	0.99	26	141	14	1

Note: Cut = 30th percentile cut point on the GRADE; Threshold = cut score identified on earlyReading measure; Concepts = Concepts of Print; AUC = Area under the Curve; Sens = Sensitivity; Spec = Specificity; Accuracy =Overall Classification Accuracy; PPV = Positive Predictive Value; NPV= Negative Predictive Value; TP = True Positives; TN = True Negative; FP = False Positives; FN = False Negatives; WB = Word Blending; WS = Word Segmenting; SW = Sight Words 150; DW = Decodable Words; NW = Nonsense Words; SR = Sentence Reading; CBM = CBMReading.

Table 11. Kindergarten Diagnostic Accuracy for fall, winter and spring using composites.

Measures	Season	Cut	Threshold	AUC	Sens	Spec	Accuracy	PPV	NPV	tp	tn	fp	fn
Full Model	Fall	30	101	0.86	0.77	0.78	0.77	0.44	0.94	32	133	40	9
O + R	Fall	30	101	0.86	0.79	0.83	0.80	0.49	0.95	34	137	36	7
C+R	Fall	30	101	0.82	0.73	0.77	0.76	0.43	0.92	30	133	40	11
LN+R	Fall	30	101	0.82	0.72	0.73	0.72	0.38	0.92	30	125	48	11
R+C+O	Fall	30	101	0.86	0.78	0.79	0.79	0.47	0.94	32	137	36	9
R+LN+O	Fall	30	101	0.86	0.79	0.83	0.80	0.49	0.95	34	137	36	7
R+LS+O	Fall	30	102	0.85	0.78	0.80	0.79	0.46	0.94	33	135	38	8
C+O+LN+R	Fall	30	102	0.86	0.76	0.85	0.78	0.46	0.96	35	132	41	6
C+O+R+WB	Fall	30	102	0.86	0.75	0.85	0.77	0.44	0.96	35	129	44	6
C+O+LS+R	Fall	30	101	0.86	0.78	0.80	0.79	0.46	0.94	33	135	38	8
Full Model	Winter	30	101	0.91	0.83	0.86	0.84	0.57	0.96	32	119	24	5
LS+R	Winter	30	101	0.86	0.82	0.81	0.82	0.54	0.94	30	116	26	7
WS+R	Winter	30	101	0.90	0.82	0.82	0.82	0.52	0.95	33	139	30	7
O+R	Winter	30	103	0.90	0.84	0.85	0.85	0.57	0.96	32	134	24	6
R+LN+WS	Winter	30	101	0.90	0.85	0.84	0.84	0.57	0.96	33	129	25	6
R+WS+O	Winter	30	101	0.91	0.82	0.84	0.83	0.54	0.95	31	132	26	7
R+LS+WS	Winter	30	102	0.90	0.80	0.87	0.81	0.52	0.96	34	123	31	5
O+LN+R+WS	Winter	30	101	0.91	0.84	0.86	0.84	0.58	0.96	32	120	23	5
LS+R+WS+WB	Winter	30	98	0.91	0.83	0.90	0.84	0.57	0.97	35	128	26	4
O+LS+R+WS	Winter	30	100	0.91	0.83	0.81	0.83	0.56	0.94	30	119	24	7

Note: Cut = 30th percentile cut point on the GRADE; Threshold = cut score identified on earlyReading measure; Concepts = Concepts of Print; AUC = Area under the Curve; Sens = Sensitivity; Spec = Specificity; Accuracy =Overall Classification Accuracy; PPV = Positive Predictive Value; NPV= Negative Predictive Value; TP = True Positives; TN = True Negative; FP = False Positives; FN = False Negatives; C = Concepts of Print; O = Onset Sounds; LN = Letter Naming; LS= Letter Sounds; R= Rhyming; WB = Word Blending; WS = Word Segmenting; SW = Sight Words 150; DW = Decodable Words; NW = Nonsense Words.

Table 11 *cont.*

Measures	Season	Cut	Threshold	AUC	Sens	Spec	Accuracy	PPV	NPV	tp	tn	fp	fn
Full Model	Spring	30	108	0.82	0.76	0.77	0.77	0.40	0.94	27	130	40	8
R+NW	Spring	30	105	0.82	0.77	0.78	0.77	0.44	0.94	31	133	40	9
R+DW	Spring	30	104	0.82	0.80	0.77	0.80	0.47	0.94	30	139	34	9
WS+DW	Spring	30	105	0.82	0.69	0.82	0.71	0.37	0.94	32	119	54	7
R+WB+NW	Spring	30	105	0.82	0.75	0.74	0.75	0.40	0.93	29	130	43	10
R+WB+DW	Spring	30	104	0.82	0.80	0.77	0.80	0.47	0.94	30	139	34	9
R+WS+DW	Spring	30	105	0.83	0.73	0.74	0.73	0.38	0.93	29	126	47	10
R+WB+NW+DW	Spring	30	104	0.82	0.76	0.74	0.76	0.39	0.93	26	129	41	9
R+WS+NW+DW	Spring	30	105	0.82	0.75	0.77	0.76	0.39	0.94	27	128	42	8
R+WB+WS+DW	Spring	30	105	0.83	0.73	0.72	0.73	0.38	0.92	28	127	46	11

Note: Cut = 30th percentile cut point on the GRADE; Threshold = cut score identified on earlyReading measure; Concepts = Concepts of Print; AUC = Area under the Curve; Sens = Sensitivity; Spec = Specificity; Accuracy =Overall Classification Accuracy; PPV = Positive Predictive Value; NPV= Negative Predictive Value; TP = True Positives; TN = True Negative; FP = False Positives; FN = False Negatives; C = Concepts of Print; O = Onset Sounds; LN = Letter Naming; LS= Letter Sounds; R= Rhyming; WB = Word Blending; WS = Word Segmenting; SW = Sight Words 150; DW = Decodable Words; NW = Nonsense Words.

Table 12. First grade Diagnostic Accuracy for fall, winter and spring using composites.

Measures	Season	Cut	Threshold	AUC	Sens	Spec	Accuracy	PPV	NPV	tp	tn	fp	fn
Full Model	Fall	30	107	0.95	0.80	0.86	0.81	0.38	0.98	18	122	30	3
WB+SR	Fall	30	107	0.94	0.80	0.86	0.81	0.38	0.98	18	122	30	3
WS+SR	Fall	30	106	0.95	0.82	0.86	0.83	0.40	0.98	18	125	27	3
WB+SW	Fall	30	93	0.93	0.86	0.81	0.85	0.44	0.97	17	130	22	4
WB+WS+SR	Fall	30	107	0.95	0.80	0.81	0.80	0.36	0.97	17	122	30	4
WB+SW+SR	Fall	30	107	0.94	0.80	0.81	0.80	0.35	0.97	17	121	31	4
WB+NW+SR	Fall	30	106	0.94	0.82	0.81	0.82	0.38	0.97	17	124	28	4
Full Model	Winter	30	104	0.99	0.94	0.96	0.94	0.73	0.99	22	124	8	1
WS+CBM	Winter	30	103	0.99	0.93	0.96	0.93	0.69	0.99	22	128	10	1
NW+CBM	Winter	30	102	0.99	0.92	0.96	0.92	0.67	0.99	22	125	11	1
WB+CBM	Winter	30	102	0.99	0.95	0.96	0.95	0.76	0.99	22	132	7	1
WS+NW+CBMR	Winter	30	101	0.99	0.96	0.96	0.96	0.81	0.99	22	129	5	1
WS+DW+CBMR	Winter	30	102	0.99	0.93	0.96	0.93	0.69	0.99	22	124	10	1
WB+NW+CBMR	Winter	30	102	0.99	0.93	0.96	0.93	0.69	0.99	22	127	10	1
Full Model	Spring	30	102	0.98	0.89	0.96	0.9	0.63	0.99	26	124	15	1
WB+CBM	Spring	30	99	0.98	0.94	0.93	0.94	0.76	0.98	25	131	8	2
WS+CBM	Spring	30	101	0.98	0.89	0.93	0.90	0.62	0.98	25	124	15	2
SW+CBM	Spring	30	100	0.99	0.91	0.96	0.92	0.68	0.99	26	128	12	1
WB+DW+CBM	Spring	30	100	0.98	0.93	0.93	0.93	0.71	0.98	25	129	10	2
WB+SW+CBM	Spring	30	100	0.98	0.93	0.93	0.93	0.71	0.98	25	129	10	2
WB+WS+CBM	Spring	30	100	0.98	0.94	0.93	0.93	0.74	0.98	25	130	9	2

Note: Cut = 30th percentile cut point on the GRADE; Threshold = cut score identified on earlyReading measure; Concepts = Concepts of Print; AUC = Area under the Curve; Sens = Sensitivity; Spec = Specificity; Accuracy =Overall Classification Accuracy; PPV = Positive Predictive Value; NPV= Negative Predictive Value; TP = True Positives; TN = True Negative; FP = False Positives; FN = False Negatives; WB = Word Blending; WS = Word Segmenting; SW = Sight Words 150; DW = Decodable Words; NW = Nonsense Words; SR = Sentence Reading; CBM = CBMReading.

Table 13. Example fitted regression models and AUC, sensitivity and specificity statistics for predicting GRADE standard scores at each time point in kindergarten grade.

	Parameter Estimates (SE)		
	Fall	Winter	Spring
	Model H	Model I	Model J
Intercept	79.33*** (2.47)	79.63*** (2.13)	62.39*** (3.62)
Onset Sounds	1.19*** (0.23)	--	--
Word Segmenting	--	0.40*** (0.06)	--
Rhyming	1.22*** (0.19)	1.50*** (0.19)	1.77*** (0.26)
Nonsense Words	--	--	7.62*** (0.94)
R^2	0.42	0.49	0.46
AUC	0.86	0.90	0.82
Sensitivity	0.79	0.82	0.77
Specificity	0.83	0.82	0.78

$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 14. Example fitted regression models and AUC, sensitivity and specificity statistics for predicting GRADE standard scores at each time point in kindergarten grade.

	Parameter Estimates (SE)		
	Fall	Winter	Spring
	Model E	Model F	Model G
Intercept	56.13*** (4.40)	27.44*** (4.54)	79.75*** (1.80)
Word Segmenting	0.39** (0.12)	--	--
Sentence Reading	13.24*** (1.03)	--	--
CBMReading	--	20.49*** (1.08)	0.31*** (0.02)
R^2	0.55	0.67	0.68
AUC	0.95	0.99	0.98
Sensitivity	0.82	0.93	0.91
Specificity	0.86	0.96	0.96

$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Figures

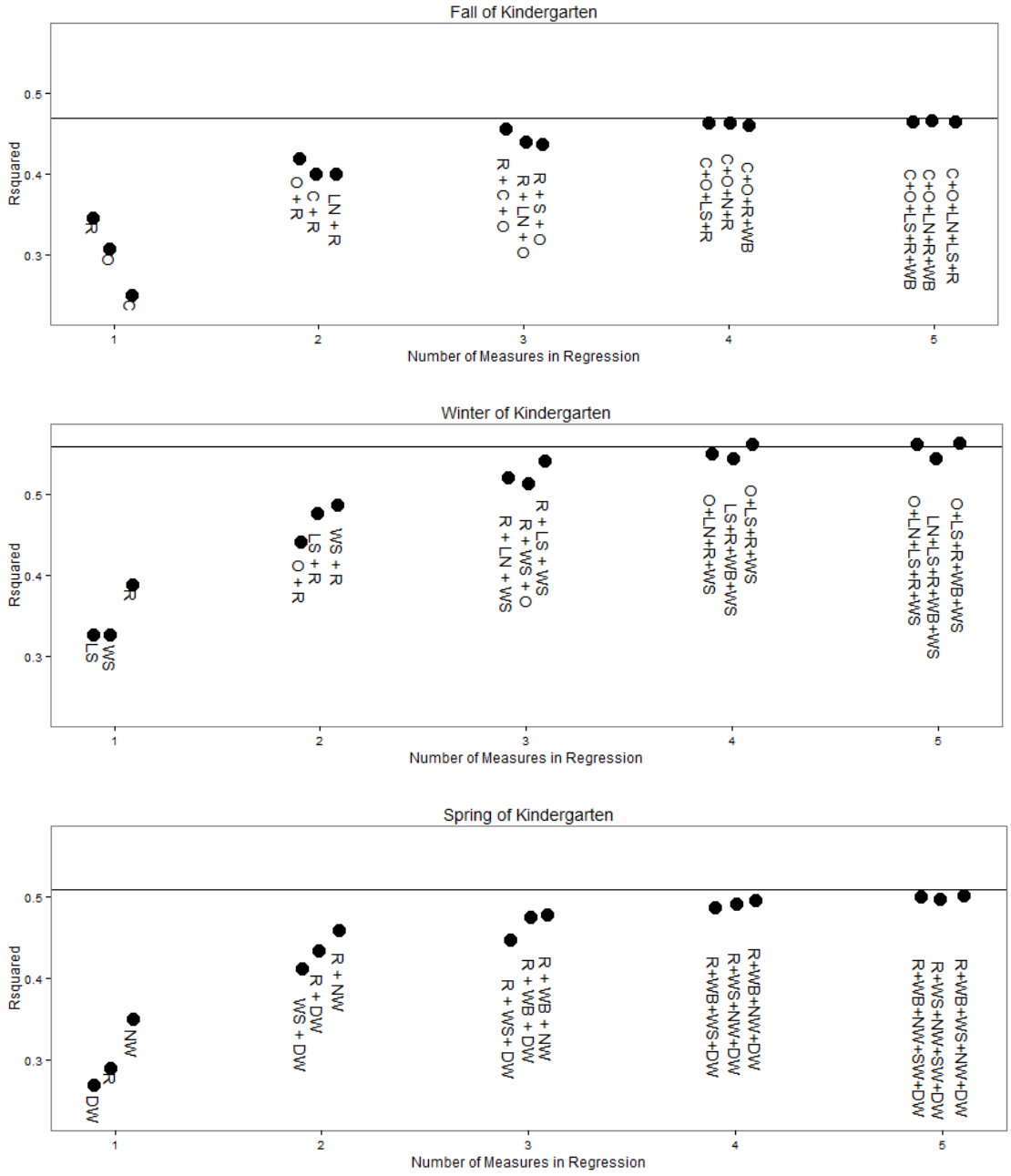

Figure 1. R^2 for top three subset regressions ranging from one to five predictors in kindergarten across time points (fall, winter and spring). Horizontal lines denote R^2 of full model. C = Concepts of Print; O = Onset Sounds; LN = Letter Names; LS = Letter Sounds; R = Rhyming; WB = Word Blending; WS = Word Segmenting; SW = Sight Words 50 ; DW = Decodable Words; NW = Nonsense Words.

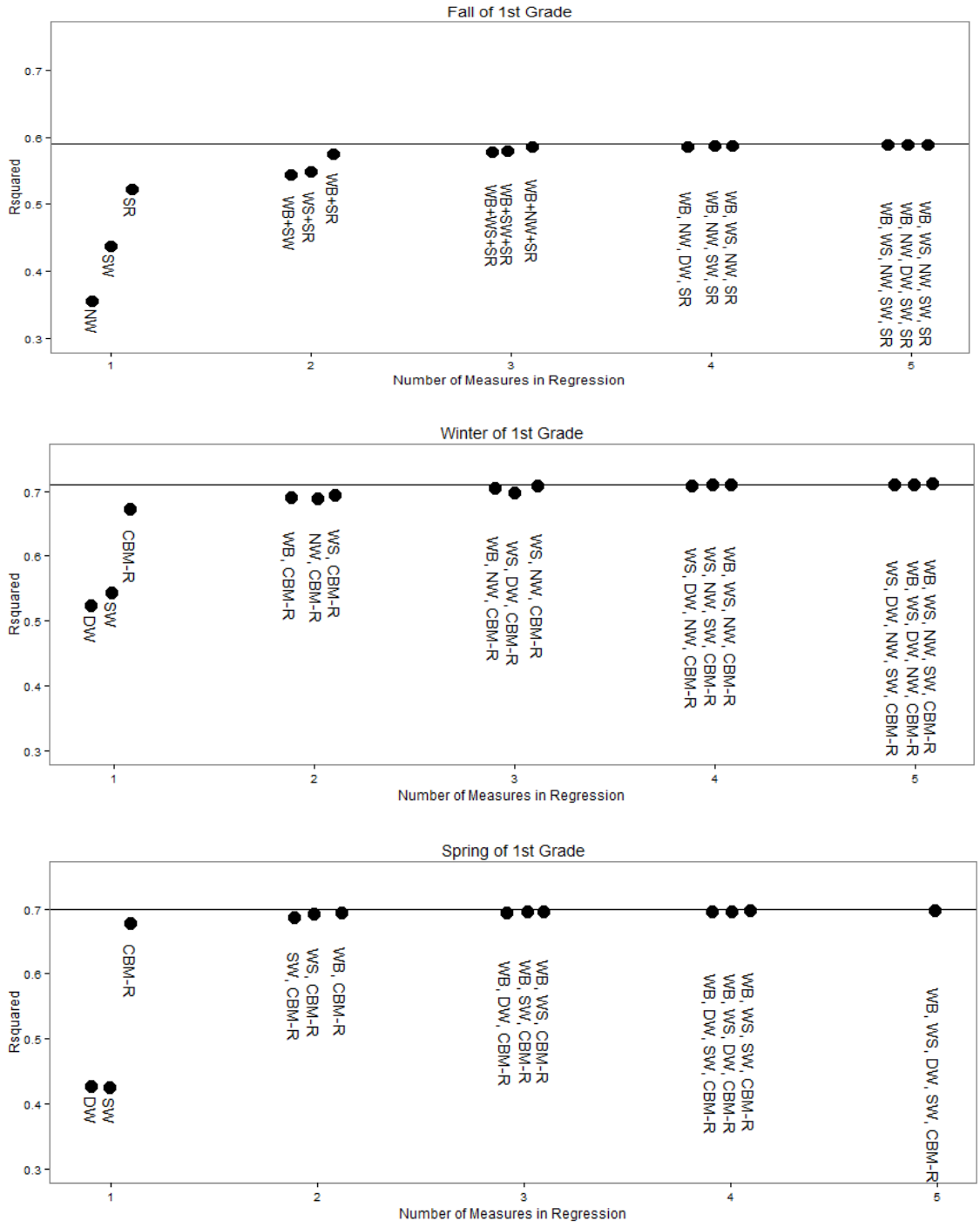

Figure 2. R^2 for top three subset regressions ranging from one to five predictors in first grade across time points (fall, winter and spring). Horizontal lines denote R^2 of full model. WB = Word Blending; WS = Word Segmenting; SW = Sight Words 150 ; DW = Decodable Words; NW = Nonsense Words; SR = Sentence Reading; CBM-R = CBMReading.

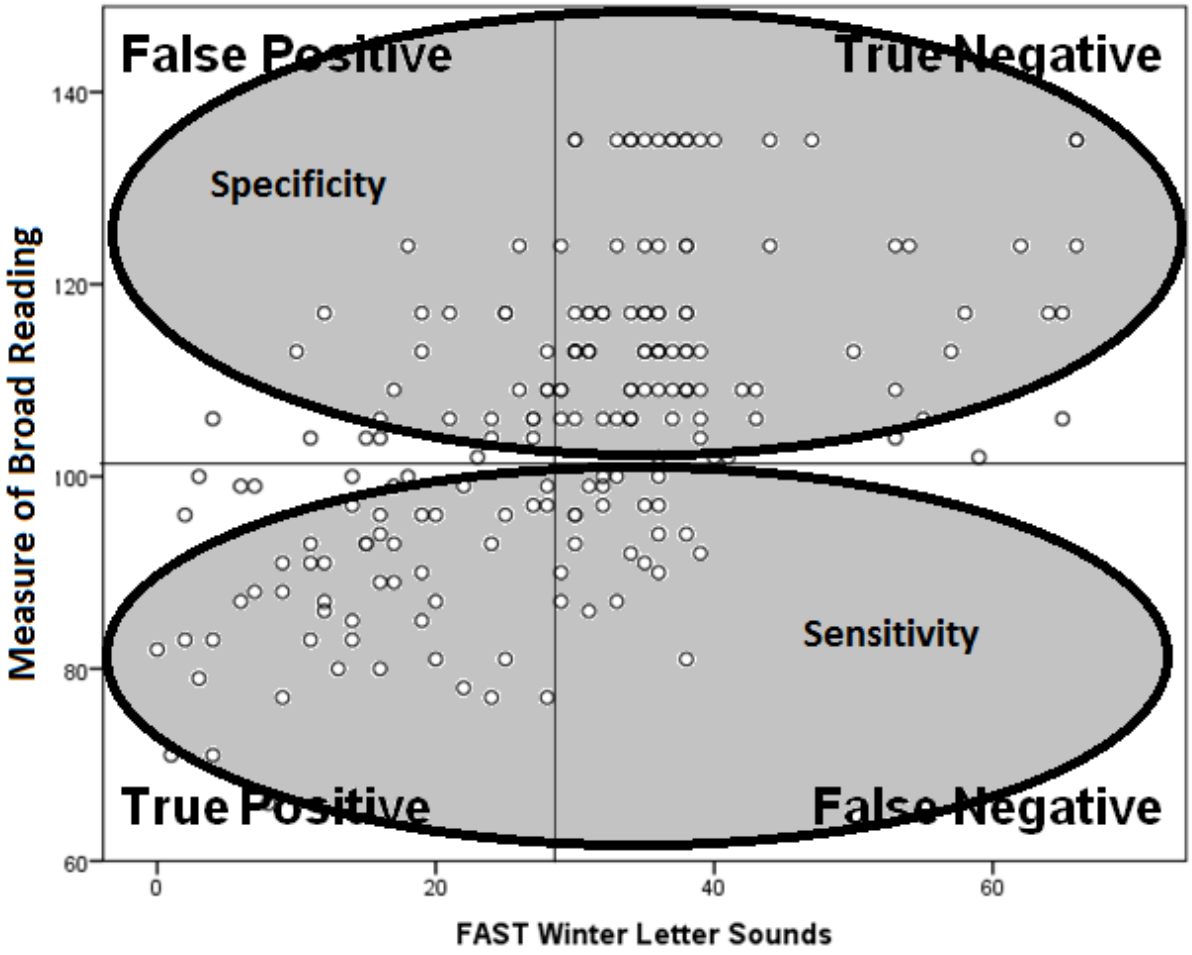

Figure 3. Example of sensitivity (TP / TP + FN) and specificity (TN / TN + FP).

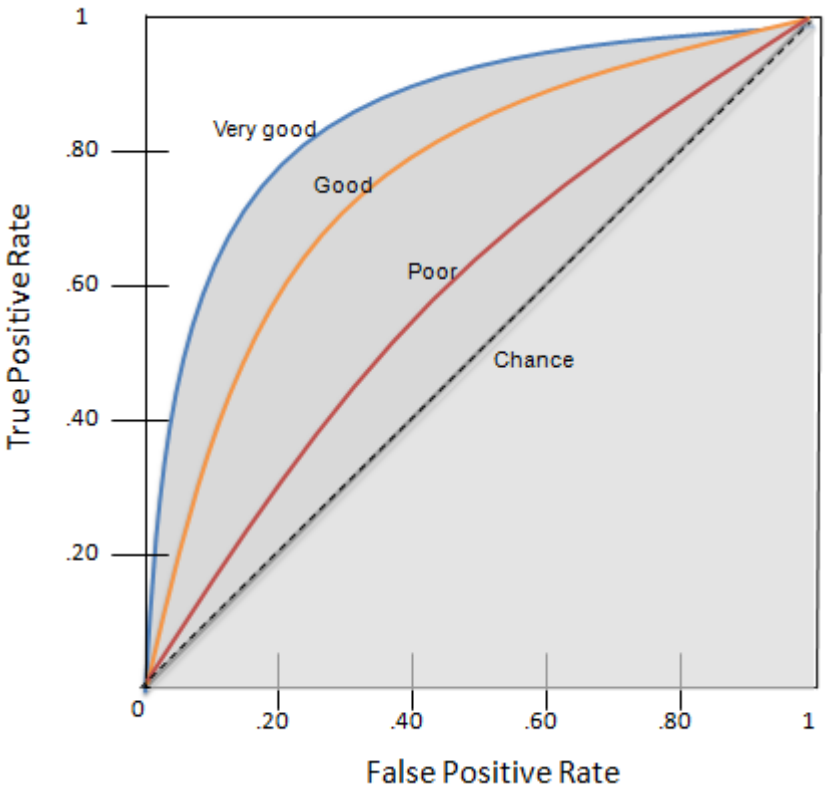

Figure 4. Example receiver operating characteristic curve with different area under the curve (AUC) values. “Very good” equals AUC of .90 or higher, “Good” equals AUC of .85 to .89, “Poor” equals AUC below .85, “Chance” equals AUC of .50.

References

Adams, M. J. (1994). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Adler, R. (Ed.). (2001). *Put Reading First: The Research Building Blocks for Teaching Children to Read*. Jessup MD: National Institute for Literacy.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational and Psychological Research Association.

Alonzo, J., & Tindal, G. (2004). District Reading Assessments, Spring 2004 Administration (Technical Report # 30). 1-66. Retrieved from http://www.brtprojects.org/publications/technical-reports

Bennett, D. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469.

Berkeley, S., Bender, W., Gregg-Peaster, L., & Saunders, L. (2009). Implementation of response to intervention. *Journal of Learning Disabilities, 42*(1), 85-95.

Blaiklock, K. (2004). The importance of letter knowledge in the relationship between phonological awareness and reading. . *Journal of Research in Reading, 27*(1), 36-57.

Bobko, P., Roth, P., & Buster, M. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*, 689-709.

Bradley, R., Danielson, L., & Doolittle, J. (2007). Responsiveness to intervention: 1997 to 2007. *Teaching Exceptional Children, 39*(5), 8-12.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (Eds.). (1984). *Classification and regression trees*. Pacific Grove, CA: Wadsworth.

Burke, M., Crowder, W., Hagan-Burke, S., & Zou, Y. (2009). A comparison of two path models for predicting reading fluency. *Remedial and Special Education, 30*(2), 84-95.

Burke, M., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32*(2), 66-77.

Burke, M., Hagan-Burke, S., Kwok, O., & Parker, R. (2009). Predictive validity of early literacy indicators from the middle of kindergarten to second grade. *The Journal of Special Education, 42*(4), 209-226.

Carnine, D., Silbert, J., Kame'enui, E., & Tarver, S. (2004). *Direct Instruction Reading* (Fourth ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Carran, D., & Scott, K. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education, 12*(2), 196-211.

Catts, H., Fey, M., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research based model and its

clinical implementation. *Language, Speech, and Hearing Services in Schools, 32*, 38-50.
Catts, H., Petscher, Y., Schatschneider, C., Sittner-Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163-`176.
Chafouleas, S., Lewandowski, L., Smith, C., Blachman, B., & 1997. (1997). Phonological Awareness Skills in Children: Examining Performance across Tasks and Ages. *Journal of Psychoeducational Assessment, 15*(334-347).
Chall, J. (1983). *Stages of reading development*. New York: McGraw-Hill.
Chall, J. (1996). *Learning to Read: The Great Debate* (3rd ed.). Orlando, Florida: Harcourt Brace & Company.
Chall, J., Roswell, F., & Blumenthal, S. (1963). Auditory blending ability: A factor in success in beginning reading. *The Reading Teacher, 17*(2), 113-118.
Chard, D., Stoolmiller, M., Harn, B., Wanzek, J., Vaughn, S., Linan-Thompson, S., & Kame'enui, E. (2008). Predicting reading success in a multilevel schoolwide reading model. *Jornal of Learning Disabilities, 41*(2), 174-188.
Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*, 55-75.
Clay, M. M. (1972). *Sand*: Heinemann Educational Books.
Clay, M. M. (1979a). *The early detection of reaind difficulties: A diagnostic survey with recovery procedures* (2nd ed.). Exeter, NH: Heinemann Educational Books.
Clay, M. M. (1979b). *Stones*: Heinemann Educational Books.
Clay, M. M. (1989). Concepts about print in english and other languages. *The Reading Teacher, 42*(4), 268-276.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum.
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (Eds.). (2003). *Applied multipe regression/correlational analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.
Compton, D., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J., Barquero, L., . . . Crouch, R. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327.
Compton, D., Fuchs, D., Fuchs, L. S., & Bryant, J. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394-409.
Cronin, V., & Carver, P. (1998). Phonological sensitivity, rapid naming, and beginning reading. *Applied Psycholinguistics, 19*, 447-461.
Cummings, K., Dewey, E., Latimer, R., & Good, R. (2011). Pathways to Word Reading and Decoding: The Roles of Automaticity and Accuracy. *School Psychology Review, 40*(2), 284-295.

Cummings, K., Kaminski, R., Good, R., & O'Neil, M. (2010). Assessing Phonemic Awareness in Preschool and Kindergarten: Development and Initial Validation of First Sound Fluency. *Assessment for Effective Intervention, 36*(2), 94-106.

Daly, E. J., III, Wright, J., Kelly, S., & Martens, B. (1997). Measures of Early Academic Skills Reliability and Validity with A First Grade Sample. *School Psychology Quarterly, 12*, 268-280.

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*(3), 184-192.

Denton, C. (2012). Response to intervention for reading difficulties in the primary grades: Some answers and lingering questions. *Journal of Learning Disabilities, 45*(3), 232-243.

Ehri, L. (1999). Phases of development in learning to read words. In J. Oakhill & R. B. (Eds.) (Eds.), *Reading development and the teaching of reading: A pscyhological perspective* (pp. 79-108). Oxford, UK: Blackwell.

Ehri, L. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188.

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills - Modified. *School Psychology Review, 30*, 33-49.

Evans, M. A., Bell, M., Shaw, D., Moretti, S., & Page, J. (2006). Letter names, letter sounds and phonological awareness: An examination of kindergarten children across letters and of letters across children. *Reading and Writing, 19*, 959-989.

Fien, H., Baker, S., Smolkowski, K., Mercier-Smith, J., Kame'enui, E., & Thomas-Beck, C. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for english learners and native english speakers. *School Psychology Review, 37*(3), 391-408.

Fien, H., Park, Y., Baker, S., Mercier-Smith, J., Stoolmiller, M., & Kame'enui, E. (2010). An examination of the relation of nonsense word fluency initial status and gains to reading outcomes for beginning readers. *School Psychology Review, 39*(4), 631-653.

Foorman, B., Francis, D., Davidson, K., Harm, M., & Griffin, J. (2004). Variability in text features in six grade 1 basal reading programs. *Scientific Studies of Reading, 8*(2), 167-197.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37-55.

Foorman, B. R., Francis, D. J., Shaywitz, S. E., Shaywitz, B. A., & Fletcher, J. M. (Eds.). (1997). *The case for early reading intervention*. Blachman, Benita A (Ed). (1997). Foundations of reading acquisition and dyslexia: Implications for early intervention.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93-99.

Fuchs, D., Fuchs, L. S., & Compton, D. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional Children, 78*(3), 263-279.

Fuchs, D., Fuchs, L. S., McMaster, K. N., & Al Otaiba, S. (2003). Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In H. L. Swanson, K. R. Harris & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 431-449). New York, NY: Guilford Press.

Fuchs, L. S., & Deno, S. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488-500.

Fuchs, L. S., Fuchs, D., & Compton, D. (2004). Monitoring Early Reading Development in First Grade: Word Identification Fluency Versus Nonsense Word Fluency. *Exceptional Children, 71*(1), 7-21.

Fuchs, L. S., Fuchs, D., Hosp, M., & Hamlett, C. (2003). The potential for diagnostic analysis within curriculum-based measurement. *Assessment for Effective Intervention, 28*, 13-22.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

Gates, A. (1926). The Gates primary reading tests. *The Teachers College Record, 28*(2), 146-178.

Gates, A. (1949). *The forty-eighth yearbook of the National Society for the Study of Education: Part II. Reading in the elementary school*. Chicago: University of Chicago Press.

Gessler-Werts, M., Lambert, M., & Carpenter, E. (2009). What special education directors say about RTI. *Learning Disability Quarterly, 32*, 245-254.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135.

Goffreda, C., Diperna, J., & Pedersen, J. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Litearcy Skills (DIBELS). *Psychology in the Schools, 46*(6), 539-552.

Good, R., & Kaminski, R. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*, 326-336.

Good, R., Kaminski, R., Dewey, E., Wallin, J., Powell-Smith, K., & Latimer, R. (2011). DIBELS Next Technical Manual. In D. Next (Ed.): University of Oregon.

Good, R. H., & Kaminski, R. A. (2002). Dynamic indicators of basic early literacy skills [6th ed.] Retrieved from http://dibels.uoregon.edu

Goodman, Y. M. (1981). Test Review: Concepts about Print Test. *The Reading Teacher, 34*(4), 445-448.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial & Special Education, 7*(1), 6-10.

Hagan-Burke, S., Burke, M., & Crowder, C. (2006). The Convergent Validity of the Dynamic Indicators of Basic Early Literacy Skills and the Test of Word Reading Efficiency for the Beginning of First Grade. *Assessment for Effective Intervention, 31*, 1-15.

Harn, B., Stoolmiller, M., & Chard, D. (2008). Measuring the Dimensions of Alphabetic Principle on the Reading Development of First Graders The Role of Automaticity and Unitization. *Journal of Learning Disabilities, 41*(2), 143-157.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological processing. *School Psychology Review, 32*(4), 541-556.

Hoffman, J., McCarthey, S., Abbott, J., Christian, C., Corman, L., Curry, C., & al., e. (1994). So what's new in the new basals? A focus on first grade. *Journal of Reading Behavior, 26*(1), 47-73.

Hoover, J., Baca, L., Wexler-Love, E., & Saenz, L. (2008). National implemention of response to intervention (RTI): Research summary. Retrieved February 23, 2014, from http://www.nasdse.org/Portals/0/NationalImplementationofRTI-ResearchSUmmary.pdf

Hoover, J., & Love, E. (2011). Supporting school-based response to intervention: A practitioner's model. *Teaching Exceptional Children, 43*(3), 40-48.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160.

Hopkins, W. (2002). A scale of magnitudes for the effect statistics. . *A review of statistics* Retrieved January 18, 2014, from http://www.sportsci.org/resource/stats/effectmag.html

Hughes, C., & Dexter, D. (2011). Response to intervention: A research-based summary. *Theory into Practice, 50*, 4-11.

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best Practices in Universal Screening. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology-V*. Bethesda, MD: National Association of School Psychologists.

Individuals with Disabilities Education Improvement Act, 20 U.S.C., Pub. L. No. 108-446 § 1400 et seq. (2004).

Indrisano, R., & Chall, J. (1995). Literacy Development. *Journal of Education 177*, 63-83.

Jenkins, J. R., Hudson, R., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582-600.

Jenkins, J. R., Peyton, J., Sanders, E., & Vadasy, P. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading, 8*(1), 53-85.

Johns, J. (1980). First graders' concepts about print. *Reading Research Quarterly, 15*(4), 529-549.

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131-140.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*(4), 174-185.

Juel, C., & Roper-Schneider, D. (1985). The Influence of Basal Readers on First Grade Reading. *Reading Research Quarterly, 20*(2), 134-152.
Kaminski, R., Cummings, K., Powell-Smith, K., & Good, R. (2008). Best Practices in Using Dynamic Indicators of Basic Early Literacy Skills for Formative Assessment and Evaluation. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology - V*. Bethesda, MD: National Association of School Psychology.
Kane, M., & Case, S. (2010). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221-240.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.
Kelly, T. L. (1927). *Interpretations of educational measures*. Yonkers, NY: Wold Book.
Kranzler, J. H., Brownell, M. T., & Miller, M. D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. *Journal of School Psychology, 36*(4), 399-415.
LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, Vol. 6*(2), 293-323.
Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100*(1), 150.
Lomax, R. G., & McGee, L. M. (1987). Young Children's Concepts about Print and Reading: Toward a Model of Word Reading Acquisition. *Reading Research Quarterly, 22*(2), 237-256.
Markell, M., & Deno, S. (1997). Effects of increasing oral reading: Generalization across reading tasks. *The Journal of Special Education, 31*(2), 233-250.
Martinez, R., & Young, A. (2011). Response to intervention: How is it practiced and perceived? *International Journal of Special Education, 26*(1), 44-52.
McAlenney, A., & Coyne, M. (2011). Identifying at-risk student for early intervention: Challenges and possible solutions. *Reading & Writing Quarterly, 27*, 306-323.
Mellard, D. F., Byrd, S. E., Johnson, E., Tollefson, J. M., & Boesche, L. (2004). Foundations and Research on Identifying Model Responsiveness-to-Intervention Sites. *Learning Disability Quarterly, 27*(4), 243-256.
Mesmer, H. (2005). Text decodability and the first-grade reader. *Reading & Writing Quarterly, 21*, 61-86.
Messick, S. (Ed.). (1989). *Validity* (Third ed.). New York: Macmillan.
Miller, A. (2002). *Subset selection in regression* (2nd ed.). New York: Chapman & Hall.
Moats. (2000). *Speech to Print: Language Essentials for Teacher*. Baltimore, MD: Paul H. Brookes Publishing Co.
Morris, D., Bloodgood, J., Lomax, R., & Perney, J. (2003). Developmental steps in learning to read: A longitudinal study in kindergarten and first grade. *Reading Research Quarterly, 38*, 302-328.

National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read* (NIH 00-4769 ed.). Washington, DC: National Institute of Child Health and Human Development.

National Research Council. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

NCRtI. (2012). Technical Standard: Validity Retrieved March 27, 2012, from http://www.rti4success.org/tools_charts/popups_screening/scoring/validity.html

Nelson, J. (2008). Beyond Correlational ANalysis of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A Classification Validity Study. *School Psychology Quarterly, 23*(4), 542-552.

No Child Left Behind Act of 2001, PL 107-110 (2001).

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*(2), 159-197.

Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing, 23*, 189-208.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184-202.

Paris, S., & Hoffman, J. (2004). Reading assessments in kindergarten through third grade: Findings from the Center for Improvement of Early Reading Achievement. *The Elementary School Journal, 105*, 199-217.

Pearson, D. P., Haertel, E., & Kamil, M. (2007). Vocabulary assessment: What we know and what we need to learn. *Research Research Quarterly, 42*(2), 282-296.

Peng, C. Y. J., Harwell, M., Liou, S. M., & Ehman, L. H. (2007). Advances in missing data methods and implications for educational research. *Real data analysis*, 31-78.

Pratt, K., Martin, M., White, M. J., Christ, T., Ardoin, S., & Eckert, T. (2011). *Development of FAIP-R Passage Sets: Level 1 (Technical Report No. 3)*. University of Minnesota, University of Georgia, Syracuse University. Minneapolis, MN.

Reschly, A., Busch, T., Betts, J., Deno, S., & Long, J. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469.

Reutzel, R., Fawson, P., Young, J., Morrison, T., & Wilcox, B. (2003). Reading enviornmental print: What is the role of concepts about print in discriminating young readers' responses? *Reading Psychology, 24*, 123-162.

Reynolds, C., & Shaywitz, S. (2009). Response to intervention: Ready or not? Or, from wait-to-fail to watch-them-fail? *School Psychology Quarterly, 24*(2), 130-145.

Rhodes, L. (1981). I can read! Predictable books as resources for reading and writing instruction. *The Reading Teacher, 36*, 511-518.

Riedel, B. (2007). The relationship between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546-567.

Ritchey, K. (2008). Assessing letter sound knowledge: A comparison of letter sound fluency and nonsense word fluency. *Exceptional Children, 74*(4), 487-506.

Rouse, H., & Fantuzzo, J. (2006). Validity of the Dynamic Indicators for Basic Early Literacy Skills as an Indicator of Early Literacy for Urban Kindergarten Children. *School Psychology Review, 35*(3), 341-355.

Schafer, J. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*, 3-15.

Schilling, S., Carlisle, J., Scott, K., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107*(5), 429-448.

Shinn, M., Good, R., Knutson, N., TIlly, W., & Collins, V. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*(3), 459-479.

Shinn, M. M., & Shinn, M. R. (2002). AIMSweb® training workbook: Administration and scoring of early literacy measures for use with AIMSweb. 1-52. Retrieved from http://aimsweb.com/index.php?page=test-of-early-literacy-samples

Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology.* (pp. 671-698). Bethesda, MD: National Association of School Psychologists.

Shinn, M. R. (2008). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology.* (pp. 243-262). Bethesda, MD: National Association of School Psychologists.

Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state mandated achievement tests: A comparions of methods. *Journal of Psychoeducational Assessment, 23*, 304-325.

Simmons, D. C., & Kame'enui, E. J. (1999). Curriculum maps: Mapping instruction to acheive instructional priorities in beginning reading kindergarten - grade 3 Retrieved from http://reading.uoregon.edu/appendices/maps.php

Spector, J. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. . *Jornal of Educational Psychology, 84*(3), 353-363.

Speece, D., Mills, C., Ritchey, K., & Hillman, E. (2003). Initial evidence that letter fluency tasks are valid indicators of early reading skill. *The Journal of Special Education, 36*, 223-233.

Speece, D., Schatschneider, C., Silverman, R., Cooper, D., & Jacobs, D. (2011). Identification of reading problems in first grade within a response to intervention framework. *The Elementary School Journal, 111*(4), 585-607.

Speece, D. L. (2005). Hitting the Moving Target Known as Reading Development: Some Thoughts on Screening Children for Secondary Interventions. [Feature Article]. *Journal of Learning Disabilities, 38*(6), 487-493.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*(4), 735-749.

Stage, S., Sheppard, J., Davidson, M., & Browning, M. (2001). Prediction of first-graders' growth in oral reading fluency using kindergarten letter fluency. *Journal of School Psychology, 39*(3), 225-237.

Stahl, S., & McKenna, M. (2001). *The concurrent development of phonological awareness, word recognition, and spelling*. Center for Improvement of Early Reading Achievement. Retrieved from http://www.ciera.org/library/archive/2001-07/200107.pdf

Stanovich, K., Cunningham, A., & Cramer, B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of Experimental Child Psychology, 39*(175-190).

Swanson, H. L., Harris, K. R., & Graham, S. (Eds.). (2003). *Handbook of learning disabilities*. New York: Guilford Press.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice, 15*(1), 55-64.

Vanderwood, M., Linklater, D., & Healy, K. (2008). Predictive accuracy of nonsense word fluency for English language learners. *School Psychology Review, 37*(1), 5.

Walton, P. (1995). Rhyming ability, phoneme identity, letter-sound knowledge, and use of orthographic analogy by prereaders. . *Jornal of Educational Psychology, 87*, 587-597.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85-120.

Williams, K. (2001). Technical Manual: Group Reading Assessment and Diagnostic Evaluation. In B. Bollard (Ed.). Circle Pines, MN.

Zirkel, P., & Thomas, L. (2010). State laws and guidelines for implementing RTI. *Teaching Exceptional Children, 43*(1), 60-73.